Nancy-Université
INPL

AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.
Il est soumis à la propriété intellectuelle de l'auteur au même titre que sa version papier. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.
D'autre part, toute contrefaçon, plagiat, reproduction illicite entraîne une poursuite pénale.

Contact SCD INPL : scdinpl@inpl-nancy.fr

LIENS

Code de la propriété intellectuelle. Articles L 122.4
Code de la propriété intellectuelle. Articles L 335.2 – L 335.10
http://www.cfcopies.com/V2/leg/leg_droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

THÈSE

présentée pour l'obtention du titre de

DOCTEUR INPL

en Procédés Biotechnologiques et Alimentaires

par Marie BRULLIARD

Infidélité de transcription et carcinogénèse.

Analyse bioinformatique et preuves de concept biologiques.

Soutenance publique prévue le 9 Juillet 2009 devant la commission d'examen :

Membres du Jury :

Rapporteurs :
M. François AMALRIC
M. Jérôme CHAILLOUX

Examinateurs :
M. François LAURENT
M. Marc PESCHANSKI
M. Luc MEJEAN (co-directeur de thèse)
M. Bernard BIHAIN (directeur de thèse)

Invités :
Mme Sandrine JACQUENET
Mme Virginie OGIER

ABRÉVIATIONS

A	adénine,
AA	acide aminé,
ADN	acide désoxyribonucléique,
ADNc	acide désoxyribonucléique complémentaire,
ARN	acide ribonucléique,
ARNm	acide ribonucléique messager,
ARNpm	acide ribonucléique pré-messager,
ARNpol II	acide ribonucléique polymérase de type II,
BLAST	basic local alignment search tool (outil d'alignement local),
Br	base de remplacement,
C	cytosine,
CCDS	consensus coding sequences (consensus des séquences codantes),
COSMIC	catalogue of somatic mutations in cancer (catalogue des mutations somatiques du cancer),
CP	canonical peptide (peptide canonique),
EJC	exon junction complex (complexe de jonction exon-exon),
EST	expressed sequence tag (étiquette de séquence exprimée),
G	guanine,
IT	infidélité de transcription,
kb	kilo base,
LBE	location based estimator (sur-estimateur du nombre de faux-positifs),
NCBI	national center of biotechnology information,
NMD	non-sense mediated decay (dégradation par l'intermédiaire des non-sens),
TIP	transcription infidelity peptide (peptide issu d'un événement d'infidélité de transcription),
PTC	premature termination codon (codon stop prématuré),
SNP	single nucleotide polymorphism (polymorphisme d'un nucleotide simple),
T	thymine,
TIAB	transcription infidelity antibody (anticorps dirigé contre un peptide issu d'infidélité de transcription),
UTR	untranslated region (région non traduite).

REMERCIEMENTS

Je tiens à remercier tout d'abord M. Bernard Bihain, qui a dirigé cette thèse avec beaucoup d'attention. Merci de m'avoir ouvert les portes de la recherche et de m'en avoir transmis la passion.
Je remercie M. Luc Méjean, co-directeur de cette thèse, pour la gentillesse qu'il manifeste à mon égard depuis de nombreuses années.

Je remercie vivement M. François Amalric, qui a fait preuve d'un réel enthousiasme dès les premières présentations et accompagne ce projet depuis quelques années.
Je remercie également M. Jérôme Chailloux, qui m'a fait découvrir, avec beaucoup de patience, les joies de la bioinformatique au cours de mon DEA.
Un grand merci à tous les deux d'avoir accepté de juger cette thèse et d'en être les rapporteurs.
Je remercie M. François Laurent et M. Marc Peschanski de me faire l'honneur de participer à ce jury.

Je remercie Sandrine Jacquenet et Virginie Ogier de me faire le plaisir de participer à ce jury. Depuis le début de cette thèse, elles sont toutes deux mes mentors en termes de rigueur scientifique. Virginie, merci d'avoir permis la conciliation délicate des données bioinformatiques et biologiques. Sandrine, merci de m'avoir enseigné l'art d'affirmer mes convictions scientifiques.

Je remercie ensuite les membres de l'équipe bioinformatique et biostatistique de Genclis, à savoir Philippe Moncuquet, Stéphanie Bolot, Valentin Harter, Olivier Collignon et Pascal Mangin. Merci pour l'ensemble de leurs travaux ; cette thèse aurait été bien plus longue et difficile sans l'aide de chacun d'entre eux.
Cette thèse présente des résultats d'études biologiques. N'ayant pas réalisé ces expériences, je tiens à citer et remercier vivement l'ensemble des personnes y ayant participé :

- Benoît Thouvenot et Lionel Bonnard, qui ont réalisé les expériences de biologie moléculaire aboutissant à la première preuve de concept de ce projet (partie 4.1.2. de ce manuscrit).

- Olivier Roitel, Virginie Ogier, Sylvianne Faron et Frances Yen, qui ont réalisé les tests immunologiques ainsi que la purification de protéines et permis ainsi de valider une seconde preuve de concept (partie 4.2.2. du manuscrit).
- Virginie Ogier, Fabrice Battais, Isabelle Sponne, Marie Barthélémy, Lionel Bonnard, Sylviane Faron, Delphine Maurice, Christelle Gonnet et Emmanuelle Guernic, qui ont travaillé à la mise au point des dosages immunologiques nécessaires à l'approche diagnostique du projet (partie 5.2.2. du manuscrit).
- Véronique Notet, qui a mis au point le modèle de cancer du poumon chez la souris, ouvrant de nouvelles perspectives (partie 5.3.3. du manuscrit).

L'ensemble des résultats présentés dans ce manuscrit représentent peu en termes de nombre de pages mais énormément en termes de travail, merci à tous.

Je conclurai ces remerciements par une note plus personnelle. Un grand merci à Charlotte et à mes Grands-parents pour leur soutien inconditionnel, ainsi qu'à mes beaux-parents pour leurs encouragements. Un merci particulier à mes parents pour m'avoir toujours accompagnée et encouragée dans mes choix, et pour m'avoir appris très tôt l'importance de l'esprit critique. Enfin, merci de tout mon cœur à Benoît pour la confiance qu'il me porte.

SOMMAIRE

1 Introduction 10

1.1 Hétérogénéité des cellules cancéreuses 10

1.1.1 Erreurs génétiques 10
1.1.2 Erreurs épigénétiques 12
1.1.2.1 La chromatine 12
1.1.2.2 Méthylation de l'ADN 13

1.2 Mesures à large échelle de l'hétérogénéité des cellules cancéreuses 15

1.2.1 Étude des mutations somatiques 15
1.2.2 Hétérogénéité du transcriptome 15
1.2.2.1 Principe d'une puce à ADN 16
1.2.2.2 Exemple d'avancée liée aux puces à ADN 17
1.2.2.3 Limites des puces à ADN 17
1.2.3 Hétérogénéité du protéome 20

1.3 Mécanismes de la transcription 20

1.3.1 Formation de l'ARN prémessager ou ARNpm 20
1.3.1.1 Le promoteur 20
1.3.1.2 Le complexe d'initiation 21
1.3.1.3 L'élongation 21
1.3.1.4 La terminaison 22
1.3.2 Maturation de l'ARN messager 22
1.3.2.1 L'addition d'une coiffe en 5' 22
1.3.2.2 L'excision et épissage 22
1.3.2.3 L'addition d'une queue polyA en 3' 23
1.3.3 Fidélité de la transcription 24
1.3.3.1 Mécanismes de surveillance 24
1.3.3.2 Mécanismes de correction 27
1.3.3.3 Exemples d'infidélité de transcription 27

1.4 Expressed Sequence Tags 30

1.4.1 Définition et mode d'obtention 30
1.4.2 Rôles et limites des ESTs 31
1.4.3 Banques de données 33

1.5 Formulation de l'hypothèse de travail 34

2 Étude préliminaire 35

2.1 Démarche bioinformatique 35

2.1.1 Extraction et tri des ESTs 35
2.1.2 Choix de 17 transcrits 35

2.1.3 Alignements 36
2.1.4 Analyse des alignements 36

2.2 Démarche statistique 39

2.2.1 Choix du test statistique 39
2.2.2 Détermination du nombre de faux positifs 41

2.3 Différence d'hétérogénéité des ESTs issues de tissus cancéreux ou normaux 42

2.4 Étude du contexte d'ADN 44

2.5 Règles de remplacement 46

2.6 Application de filtres bioinformatiques 49

2.6.1 Filtre des ESTs chimériques, des homologues et des pseudogènes 49
2.6.2 Filtre des extrémités des alignements 49
2.6.3 Normalisation des longueurs des ESTs 50
2.6.4 ESTs issues de lignées 51

3 Extension du procédé d'analyse bioinformatique au transcriptome entier 53

3.1 Données et outils utilisés 53

3.1.1 Optimisation de la démarche bioinformatique 53
3.1.2 Les séquences de référence ARN 55
3.1.3 Mise à jour des ESTs utilisées 55
3.1.4 Mise à jour du BLAST 55
3.1.5 Différents évènements d'infidélité de transcription 55
3.1.6 Mise en place d'un second test statistique 56

3.2 Filtres appliqués 57

3.2.1 Filtre d'alignements 57
3.2.2 Filtre de positions 57
3.2.2.1 Substitutions 58
3.2.2.2 Délétions 58
3.2.2.3 Insertions 59

3.3 Application des tests statistiques 60

3.4 Résultats 62

3.4.1 Évènements simples 62
3.4.1.1 Effectifs 62
3.4.1.2 Résultats des tests statistiques 64
3.4.2 Évènements multiples 67

3.4.2.1 Modification des paramètres du MegaBLAST 67
3.4.2.2 Résultats 68

3.5 Étude du contexte d'ADN 70

3.5.1 Évènements d'infidélité de transcription 70
3.5.1.1 Règles de substitutions 70
3.5.1.2 Contexte n-uplet des délétions 77
3.5.1.3 Contexte des insertions 78
3.5.2 Mutations somatiques 80
3.5.2.1 Étude de tumeurs du sein et du colon 81
3.5.2.2 Cas particulier des transitions C : G → A : T 82
3.5.2.3 Analyse du contexte d'ADN 82

4 Prédictions bioinformatiques et résultats biologiques 89

4.1 Détection d'un ARNm présentant une délétion à une position prédite par la bioinformatique 89

4.1.1 Prédiction bioinformatique 89
4.1.2 Preuve de concept 89
4.1.3 Conclusion 91

4.2 Infidélité de transcription affectant le codon stop 92

4.2.1 Définition et nature des stops alternatifs 92
4.2.2 Preuve de concept 94

4.3 Conséquences d'une délétion sur la protéine et prédiction de peptides issus d'infidélité de transcription 97

4.3.1 Impact codant d'une délétion 97
4.3.2 Prédiction des TIPs (Transcription Infidelity Peptides) 99
4.3.3 Formulation de l'hypothèse biologique 99

5 Utilisation des TIPs pour le diagnostic des cancers 101

5.1 Détection d'anticorps dirigés contre les TIPs 101

5.1.1 TIPs sélectionnés pour la validation biologique 101
5.1.2 Principe du test 103
5.1.3 Mise en place de contrôles négatifs 103
5.1.3.1 Choix des contrôles négatifs 103
5.1.3.2 Test des contrôles négatifs 104

5.2 Discrimination sérum cancéreux / sérum non cancéreux 104

5.2.1 Données cliniques 105
5.2.2 Données brutes 105
5.2.3 Analyse statistique des résultats 106

5.3 Perspectives d'étude chez Mus musculus 109
5.3.1 Analyse des ESTs de *Mus musculus* 109
5.3.2 Identification d'évènements homologues 112
5.3.3 Preuve de concept *in vivo* 114

6 Conclusions et perspectives 116

1 Introduction

Le cancer est une maladie chronique résultant de dérèglements du fonctionnement des cellules vieillissantes.

Le rapport de l'IARC (International Agency for Research on Cancer) de l'année 2008 estime à 12,4 millions l'incidence ou morbidité (*i.e.* le nombre de nouveaux cas de cancer détectés dans le monde en 2008) et à 7,8 millions la mortalité (*i.e.* le nombre de décès liés au cancer). Le type de cancer ayant la plus grande incidence est le cancer du poumon chez l'homme, avec 960.000 nouveaux cas par an et 850.000 décès, et le cancer du sein chez la femme, avec 1,1 millions de nouveaux cas par an et 410.000 décès.

1.1 Hétérogénéité des cellules cancéreuses

Le cancer peut présenter des formes et des évolutions variables d'un patient à l'autre. Dès le début des années 1980, la tumeur est décrite comme un ensemble constitué de sous-populations de cellules hétérogènes (du point de vue morphologique, caryotypique ou encore histologique) et fonctionnellement différentes [1,2]. Les cellules subissent en effet des modifications métaboliques et comportementales les conduisant à proliférer de manière excessive, à échapper à la surveillance du système immunitaire et à envahir des tissus plus éloignés [3].

1.1.1 Erreurs génétiques

Boveri, en 1914, avait noté que les structures nucléaires des cellules tumorales malignes diffèrent très fréquemment de celles des cellules normales et supposé qu'elles étaient la marque de modifications du patrimoine héréditaire survenues dans une cellule au sein d'un tissu et la cause des aberrations de comportement des cellules malignes. Finalement, l'hypothèse de l'origine nucléaire des cancers de Boveri a été reformulée dans les années 1980 pour devenir l'hypothèse des mutations somatiques comme origine des cancers (SMT Somatic Mutation Theory).

Le cancer est défini aujourd'hui comme une maladie génétique résultant de l'accumulation de mutations ou altérations génétiques dans les oncogènes et les gènes suppresseurs de tumeurs [4]. Les oncogènes (*e.g.*, KRAS, v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog [5]) sont des gènes qui, lorsqu'ils sont mutés, activent la prolifération cellulaire [6]. Les gènes suppresseurs de tumeur (*e.g.*, TP53, tumor protein p53 [7]) sont des gènes qui, une fois inactivés

par une mutation, contribuent au développement tumoral [8]. Les mutations peuvent affecter une ou plusieurs bases. De grandes régions d'ADN, contenant plusieurs gènes, peuvent en effet être éliminées ou transloquées à un autre locus génomique [9].

Parmi les gènes impliqués, certains sont fortement liés à un type de cancer donné : BRCA1 (breast cancer 1, early onset), par exemple, est muté dans 50 à 60% des cancers du sein [10,11]. D'autres gènes au contraire sont mutés de façon récurrente dans différents types de cancer, comme TP53 qui est muté dans plus de 50% des cancers, tous types confondus [12,13]. Il reste toutefois difficile d'établir le nombre minimal d'altérations génétiques nécessaires à l'émergence d'une transformation maligne dans une cellule.

Trois courants sont proposés :

- ✓ Le premier propose comme hypothèse une accumulation de mutations au sein de gènes associés au cancer, accumulation qui se fait de manière séquentielle et uniforme, chaque mutation conférant un avantage sélectif [14].

  Une étude récente s'est concentrée sur le séquençage de quelques tumeurs, permettant ainsi d'appréhender l'importance des mutations somatiques au niveau d'une tumeur. Le séquençage de régions codantes et de jonctions intron-exon d'ADN génomique extrait de 11 tumeurs du sein et 11 tumeurs colorectales a ainsi permis l'identification de 1307 mutations somatiques ayant un impact codant [15]. Les gènes affectés dépendent du type de tumeur et sont même spécifiques de chaque tumeur ; ce résultat établit clairement l'hétérogénéité des mutations somatiques impliquées dans le cancer. Notons que les mutations somatiques restent des événements rares (3,1 par 10^6 bases), générant en moyenne 90 substitutions d'AA par tumeur [15].

- ✓ Le second courant repose sur l'hypothèse qu'un cancer se développe par vagues successives de sélections clonales [16]. Les mutations dites clonales sont sélectionnées par la tumeur et présentes dans la plupart des cellules cancéreuses. Cette hypothèse implique l'existence d'un nombre limité de gènes impliqués dans la plupart des cancers.

- ✓ Le troisième courant est l'hypothèse du phénotype mutateur, dans laquelle des mutations somatiques dites aléatoires, plus fréquentes mais distribuées aléatoirement et non sélectionnées sont présentes sur seulement quelques cellules de la tumeur [17-19]. Ainsi, les génotypes de toutes les cellules d'une même tumeur ne seraient pas identiques. Notons que les techniques classiques de séquençage ne détectent une mutation somatique que si elle affecte au moins 10% des cellules tumorales [20]. Ainsi, pour accéder aux mutations somatiques aléatoires, il est nécessaire de descendre à l'échelle d'une molécule

d'ADN. Bielas *et al.* [21] ont mis au point une telle technique, appelée «random mutation capture». Ils montrent ainsi que le taux de mutations somatiques dans les tissus humains est de moins de 1 mutation par 10^8 nucléotides dans un tissu normal et de l'ordre de 100 à 500 mutations par 10^8 nucléotides selon le type de tumeur étudiée. Ainsi, même les mutations somatiques aléatoires restent des évènements rares. Par ailleurs, seules les mutations non silencieuses ont un impact sur la séquence protéique, limitant ainsi l'hétérogénéité générée. Enfin, il est admis que seules 30% des mutations non silencieuses modifient la fonction protéique [22].

1.1.2 Erreurs épigénétiques

L'épigénétique est l'étude des changements héréditaires dans la fonction des gènes, ayant lieu sans altération de la séquence d'ADN [23,24]. Toutes les cellules possèdent le même matériel génétique, mais seuls certains gènes sont exprimés en fonction du type cellulaire. L'information sur le verrouillage ou le déverrouillage de certaines portions du génome est portée par des facteurs épigénétiques. D'autre part, la réparation de l'ADN suite à une mutation ne peut s'effectuer qu'à la condition que l'ADN soit accessible, ce qui suppose la désorganisation de la chromatine, puis le rétablissement de son organisation initiale.

1.1.2.1 La chromatine

La chromatine est la forme sous laquelle se présente l'ADN dans le noyau et correspond à l'association entre l'ADN et les histones, qui sont des protéines de structure.

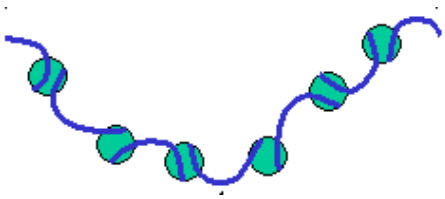


Figure 1 : brin de chromatine.

L'ADN est représenté par un trait bleu et les histones par des « perles » vertes.

Le niveau de transcription est directement lié au niveau de condensation de l'ADN. En effet, les gènes ne peuvent être transcrits que si la chromatine est décompactée.

Les 4 types de modifications que peuvent subir les histones sont l'acétylation, la méthylation, la phosphorylation et l'ubiquitination [25,26]. L'acétylation des histones ainsi que la méthylation des histones ou de l'ADN sont les modifications post-traductionnelles les plus directement liées à l'état transcriptionnel de la région d'ADN impliquée.

L'acétylation est le transfert d'un groupement acétyle provenant de l'acétyle-coenzyme A sur le groupement ε- amino de certains résidus lysine. L'acétylation des histones produit une chromatine plus flexible, où l'ADN est plus accessible.

1.1.2.2 Méthylation de l'ADN

La méthylation de l'ADN se produit principalement dans des régions appelées îlots CpG [27]. Les îlots CpG sont des régions génomiques de 0,5 à plusieurs milliers de kb contenant 60-70% de dinucléotides CpG [28-30]. Le génome humain en compte près de 30.000 [31]. La méthylation des cytosines entraîne une modification de l'architecture de la fibre de chromatine qui aboutit à un compactage des nucléosomes, empêchant l'accès des facteurs de transcription aux sites de fixation situés dans le promoteur. Des méthylations accidentelles ou aberrantes ont été décrites dans différents types de tumeurs au niveau du promoteur ou du premier exon de différents gènes impliqués dans la carcinogenèse [32,33]. Une hyperméthylation située au niveau du promoteur ou du premier exon conduit à un arrêt de l'expression (impact particulièrement important sur les gènes suppresseurs de tumeur), alors qu'une hypométhylation engendrera une activation de la transcription (impact sur les oncogènes).
La méthylation de l'ADN génomique est contrôlée par les ADN méthyltransférases (DNMT). L'expression des DNMT est augmentée dans différents types de cancers : sein [34], prostate [35], colon [36], endomètre [37]. L'augmentation de la synthèse de ces protéines est fortement associée à l'hyperméthylation des îlots CpG situés dans les régions promotrices de différents gènes suppresseurs de tumeurs. Ce phénomène est associé à l'absence des protéines codées par les gènes suppresseurs de tumeurs dans les cellules cancéreuses et au développement du phénotype malin.

Gènes suppresseurs de tumeur	Fonction	Type de cancer associé à cette hyperméthylation	référence
AKAP12	Transduction du signal	Estomac	Choi, 2004
APC	Prolifération cellulaire, migration et adhésion	Colon	Hiltunen, 1997
BRCA1	Réparation de l'ADN	Sein, ovaire	Esteller, 2000
CASP8	Apoptose	Neuroblastome	Casciano,2004
CDH1	Adhésion cellule-cellule	Sein, prostate,	Darwanto,2003,

		colon	Graff, 1995
CDKN2A	Inhibiteur de kinases cycline dépendant	Lymphome	Herman, 1995
CDKN2B		Leucémie	Melki, 1999
DAPK1	Apoptose induite par l'interferon	Poumon, lymphome	Kastenellenbogen, 1999
GSTP1	Empêche l'oxydation de l'ADN	Prostate	Lee, 1994
ING1	Croissance cellulaire et apoptose	Poumon	Kameyama, 2003
KISS1	Chimiotactisme et invasion	Sein	Stark, 2005
LATS2	transcription médiée par les récepteurs d'androgène	Sein	Takahashi, 2005
MLH1	Réparation de l'ADN	Colon	Veigl, 1998
PTEN	Régulation négative de la voie AKT/PKB	Colon	Goel, 2004
RASSF1	Arrêt du cycle cellulaire	Rein	Morissey, 2001
RB1	Répression de la transcription des gènes cellulaires	Rétinoblastome	Stirzaker, 1997
SMAD4	Prolifération cellulaire	Colon	Ando, 2005
SOCS1	Contrôle négatif de la voie des cytokines	Pancréas	Fukuhima, 2003
STK11	Transduction du signal	Poumon	Sanchez-cespedes, 2002
TIMP3	Inhibition de matrices métalloprotéinases	Colon, rein, cerveau	Bachman, 1999
TP53	Régulation du cycle cellulaire	Leucémie	Agirre, 2003
WWOX	Régulation de la transcription, dégradation des protéines	Poumon, sein, vessie	Iliopoulos, 2005
VHL	Activation de l'angiogénèse	rein	Morissey, 2001

Table 1 : Liste de gènes suppresseurs de tumeurs soumis à une hyperméthylation tirée de Luczak et al. [33].

Ainsi, les facteurs génétiques et épigénétiques contribuent au développement tumoral et sont source d'hétérogénéité dans les tissus cancéreux. Cette hétérogénéité est approchée par différentes techniques. Les mutations somatiques sont recherchées par séquençage direct de l'ADN. La perturbation du niveau d'expression des gènes est caractérisée par des études de transcriptomique. Enfin, la production de protéines issues de gènes mutés est mesurée par des études de protéomique.

1.2 Mesures à large échelle de l'hétérogénéité des cellules cancéreuses

1.2.1 Étude des mutations somatiques

Le *Cancer Gene Census* décrit 347 gènes humains porteurs de mutations somatiques [38]. COSMIC (Catalogue Of Somatic Mutations In Cancer) est une base de données constituée des mutations somatiques des gènes connus pour être impliqués dans la physiopathologie des cancers [39]. La version 40 de COSMIC fournit quelques 50.000 mutations somatiques détectées après analyse de 4.800 gènes. COSMIC combine les résultats issus de la littérature avec les données de séquençage du Sanger Institute (Cancer Genome Project). Il est intéressant de noter qu'il existe une base de données dédiée aux mutations somatiques du gène TP53 (http://www-p53.iard.fr/), qui référence 24.000 mutations.

1.2.2 Hétérogénéité du transcriptome

Une cellule différenciée est caractérisée par l'expression de 10.000 à 15.000 gènes différents. La quantité produite de chaque ARNm varie de quelques copies à plusieurs dizaines de milliers de copies. Chaque type cellulaire possède donc un profil d'expression d'ARN, ou transcriptome, qui représente une information relative à l'état physiopathologique de la cellule à un instant donné, constituant ainsi une signature moléculaire. La transcriptomique, ou génomique fonctionnelle, correspond à l'étude de l'expression des gènes par des techniques incluant la PCR quantitative, la méthode SAGE, ou encore les puces à ADN.

Les puces à ADN permettent d'étudier l'expression de milliers de gènes simultanément. Pour avoir un ordre d'idées du nombre d'études réalisées depuis leur première application en 1995 [40], une simple recherche d'articles portant sur le cancer humain sur le site d'Affymetrix (https://www.affymetrix.com/publications) donne 1.267 réponses au 31 décembre 2008.

1.2.2.1 Principe d'une puce à ADN

Le principe d'une puce à ADN, ou biopuce, repose sur la capacité d'hybridation d'un ADN avec une sonde complémentaire.

Une puce à ADN est constituée de plusieurs milliers de fragments d'ADN ou sondes fixées à un support (verre, silicium, plastique). Les ARN totaux extraits de cellules ou tissus sont amplifiés, puis les ARNm sont rétrotranscrits en ADNc et enfin marqués (radioactivité ou fluorescence). A noter que l'intérêt de l'utilisation de fluorochromes (Cyanine 3 (vert) et Cyanine 5 (rouge)) est de tester deux conditions différentes sur une même puce. Seuls les ADNc complémentaires de ceux présents sur la puce vont s'apparier ; ils seront ensuite détectés grâce à leur marquage.

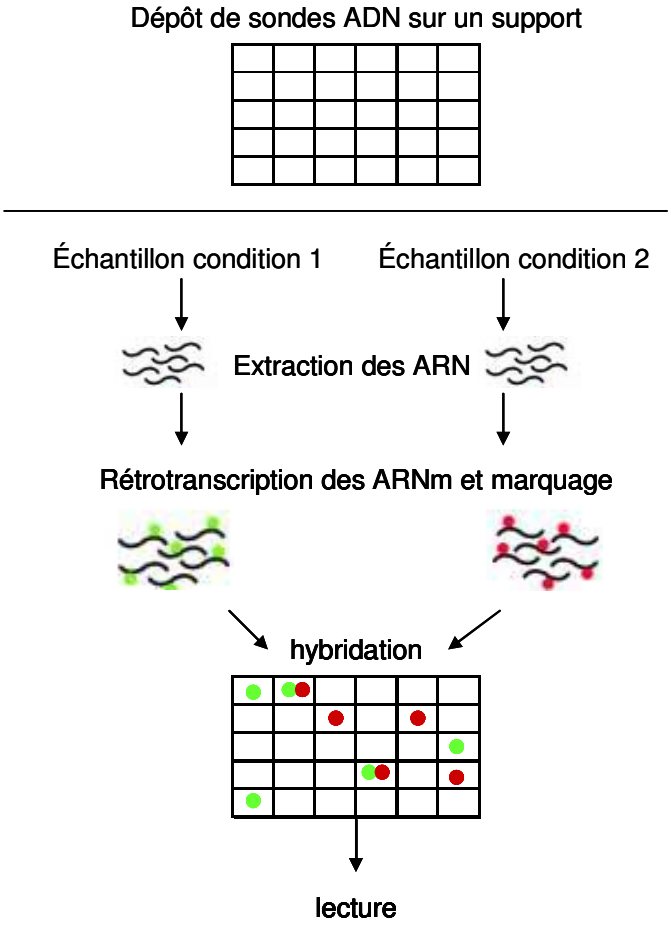


Figure 2 : principe des puces à ADN.

L'analyse des résultats se fait par calcul du ratio de fluorescence rouge/fluorescence verte pour chacun des spots, ce ratio permet donc de visualiser l'expression différentielle des gènes dans les deux échantillons.

1.2.2.2 Exemple d'avancée liée aux puces à ADN

De nombreuses études ont été réalisées pour étudier l'expression des gènes dans les différents types et sous-types de cancers. Les avancées liées aux études transcriptomiques sont indiscutables ; l'exemple détaillé ici sera celui du cancer du sein. Les premières études ont permis de confirmer l'importance du statut des récepteurs d'œstrogène (ER), distinguant les tumeurs ER négatives (*basal-like*) et les tumeurs ER positives (*luminal-like* surexprimant l'oncogène HER2 ou ERBB2, v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)). Dans le cadre du cancer du sein, une chimiothérapie ou un traitement hormonal permettent de réduire le risque de métastase de 30%. Ne sachant pas à l'avance si la tumeur risque ou non d'envahir d'autres tissus, toutes les patientes subissent un lourd et pénible traitement alors que seuls 20-30% en tirent bénéfice. Une équipe de chercheurs hollandais a étudié 295 patientes atteintes de tumeurs primaires du sein et identifié une signature moléculaire de 70 gènes capables d'identifier une sous-population de patientes susceptibles de rechuter ou de développer des métastases avant 5 ans [41,42]. Le premier groupe de 180 patientes classées avec un mauvais pronostic présentait un taux de survie à 10 ans de 54,6% (±4,4). Le second groupe de 115 patientes classées avec un bon pronostic présentait un taux de survie à 10 ans de 94,5% (±2,6). Ce profil d'expression s'est avéré meilleur que les critères cliniques ou histologiques précédemment établis. Cette étude a ensuite été confirmée dans une seconde population de 123 patientes : 48% ont été classées avec un mauvais pronostic et présentaient une survie à 5 ans de 82% (±5) ; 52% ont été classées avec un bon pronostic et présentaient une survie à 5 ans de 97% (±2) [43]. Une étude indépendante portant sur 307 patientes issues de 5 centres de recrutement européens a également confirmé l'intérêt de cette signature moléculaire de 70 gènes [44]. Par ailleurs, un autre groupe [45] a mis en évidence une signature de 76 gènes, elle aussi confirmée par une seconde étude [46]. L'intersection des deux ensembles contient peu de gènes, ce qui semble indiquer qu'il existe différents types de profils d'expression liés à la même pathologie. Néanmoins, une étude des deux ensembles de gènes testés sur 198 patientes a montré que les pronostics prédits étaient identiques dans 71% des cas [47].

1.2.2.3 Limites des puces à ADN

Exemples

Deux études visant à prédire la survie de patients atteints de lymphome à cellules B et ayant reçu une chimiothérapie montrent des résultats totalement divergents [48,49]. La première donne

comme résultat une combinaison de 17 gènes et la seconde une combinaison de 13 gènes. L'intersection des deux ensembles est vide.

L'exemple détaillé dans le paragraphe ci-dessus montre l'existence de deux études récentes portant sur le cancer du sein et donnant des résultats cohérents mais reposant sur des signatures moléculaires très différentes [42,45]. Ainsi, deux analyses transcriptomiques différentes fournissent des combinaisons de gènes totalement différentes.

Pistes d'explication

Il s'agit là d'exemples isolés d'études ayant pu être reproduites. De nombreuses critiques sont en effet émises sur le manque général de reproductibilité des études de transcriptomique [50-53].

En 2003, Tan *et al.* [50] publient une étude dans laquelle le niveau d'expression des ARNm issus de la même culture cellulaire (cellules PANC-1 en milieu enrichi en sérum et 24h après élimination du sérum) est testé sur 3 puces différentes (Affymetrix, Agilent et Amersham). Selon la technologie utilisée, la sonde spécifique d'un gène d'intérêt varie ; en d'autres termes, chaque fournisseur possède des sondes spécifiques. Le diagramme ci-dessous montre le nombre de gènes identifiés comme étant différentiellement exprimés dans les 3 expériences :

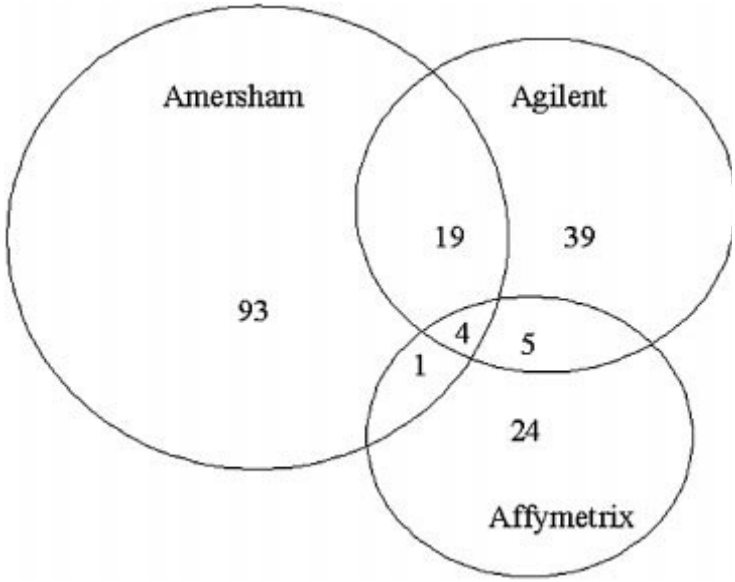


Figure 3 : nombre de gènes différentiellement exprimés dans 3 puces différentes, tiré de [50].

Cam et son équipe concluent à la nécessité d'accroître la maîtrise technologique afin de pouvoir comparer les résultats issus de puces venant de fournisseurs différents. Des investigations plus poussées ont montré qu'un nombre important de sondes jusqu'alors utilisées étaient incorrectes [54,55] et qu'un tiers d'entre elles n'étaient pas présentes dans la base

RefSeq du NCBI [56,57]. D'autre part, il apparaît que les sondes synthétisées directement sur le support peuvent contenir des erreurs. Par ailleurs, les sondes des différents fournisseurs pouvaient correspondre à des régions différentes du même gène et en particulier des régions soumises à l'épissage. Ce constat a entraîné la production de sondes spécifiques des variants d'épissage et non des gènes. Cet élément ne suffit pas à lui seul à expliquer les différences de résultats.

Les sondes utilisées sur les puces à ADN peuvent être de différentes natures (oligonucléotides de 25-30 ou 60-70 bases ou produits de PCR pouvant aller jusqu'à quelques milliers de bases). En outre, les techniques d'application des sondes sur le support sont diverses, ainsi que les techniques d'acquisition et de traitement des données [58]. En travaillant sur les puces Affymetrix GeneChip U95A/Av2, Zhang et al. ont montré que 10% des sondes étaient non spécifiques, *i.e* correspondaient à plusieurs gènes différents ; il peut donc y avoir hybridation croisée pour 10% des sondes de cette puce [59]. De plus, 10% des sondes ne sont pas fonctionnelles dans la mesure où le transcrit correspondant, même s'il est présent, ne s'hybride pas. Le même travail réalisé sur les puces U133A fournit des résultats similaires [59].
Il faut souligner le fait que l'expression des transcrits dits rares n'est pas mesurée avec la même sensibilité que l'expression des transcrits abondants. Chez le rat, 10 copies d'ARNm par cellule semble être le minimum requis pour obtenir une détection reproductible [60]. Néanmoins, la reproductibilité des variations d'expression des transcrits rares, comme les facteurs de transcription, reste insuffisante [61,62].
La reproductibilité des résultats ne garantit pas leur fiabilité. En effet, différentes études montrent que les résultats de puces à ADN (sondes oligonucléotidiques) sont reproductibles : une sonde donnée s'hybride au même nombre de transcrits marqués dans des expériences répétées [63,64]. Mais la reproductibilité, bien que nécessaire, n'est pas suffisante. Une signature de gènes établie sur une population doit pouvoir s'appliquer à une autre population présentant les mêmes caractéristiques. Ceci n'est actuellement pas le cas, et il apparaît également nécessaire, avant d'incriminer la méthode elle-même, de s'intéresser aux populations d'étude, à savoir les conditions de prélèvement des échantillons, les conditions de congélation ou encore les conditions dans lesquelles l'ARN a été stabilisé.

Il existe donc de nombreux problèmes techniques et probablement conceptuels liés au manque de cohérence des résultats issus des études de transcriptomique. Soulignons également que les difficultés de convergence peuvent être liées à l'hétérogénéité inhérente à la maladie. En effet,

un transcriptome en perpétuel changement et régi par des règles encore méconnues reste une explication plausible du manque relatif de fiabilité des données obtenues par puce.

1.2.3 Hétérogénéité du protéome

Un gène dysfonctionnel peut avoir pour conséquence une dérégulation du niveau de production des protéines, la production de protéines altérées ou l'absence de production de protéines.

Là encore, les études de protéomique se sont multipliées ces dernières années, mais comme pour la transcriptomique, le manque de convergence des résultats expérimentaux est criant. Ce travail se focalisant sur l'origine de l'hétérogénéité des transcrits cancéreux, l'hétérogénéité du protéome tumoral ne sera pas développée ici.

Les cellules cancéreuses sont donc hétérogènes. Les différentes mesures de cette hétérogénéité à l'échelle du transcriptome et du protéome ne sont pas satisfaisantes et ne permettent pas de caractériser avec précision et assurance l'état cancéreux. Nous nous sommes focalisés sur une nouvelle source d'hétérogénéité détectée au niveau de la séquence de l'ARN et en l'absence de variation de la séquence d'ADN. Nous allons donc examiner les mécanismes de transfert de l'information de l'ADN à l'ARN, à savoir la transcription.

1.3 Mécanismes de la transcription

La transcription est le processus de copie de l'information portée par l'ADN en ARN. L'expression peut être régulée selon le stade de développement, le type cellulaire, l'environnement… Chez les eucaryotes, trois ARN polymérases (ARNpol) assurent la transcription : l'ARNpol I pour les ARN ribosomiques ou ARNr (28S, 18S et 5,8S), l'ARNpol II pour les ARN messagers ou ARNm, et l'ARNpol III pour les petits ARN (ARN de transfert, ARNr 5S, petits ARN nucléaires). Nous nous contenterons ici de décrire brièvement la formation d'ARNm.

1.3.1 Formation de l'ARN prémessager ou ARNpm

1.3.1.1 Le promoteur

Le promoteur correspond à une région non transcrite de l'ADN, située en amont de la région transcrite. La séquence du promoteur permet le recrutement de l'ARNpol II. Certaines parties

du promoteur ou "boîtes" ont une importance particulière parce qu'elles sont reconnues spécifiquement par différentes protéines du complexe d'initiation (voir paragraphe suivant) :

- ✓ La plus étudiée est la "boîte TATA" riche en thymine et adénine. Elle est généralement située vers -25 à -30 nucléotides du site de démarrage de la transcription,
- ✓ des éléments proximaux :
  - la "boîte CAAT" est facultative et contient de la cytosine. Elle est située vers -120 à -80 nucléotides du site de démarrage de la transcription,
  - la "boîte GC" est facultative également et est riche en guanine et cytosine. Quand elle existe, elle est localisée entre la boîte CAAT et la boîte TATA.

1.3.1.2 Le complexe d'initiation

Chez les eucaryotes, le promoteur est reconnu par l'ARNpol II avec de nombreux co-facteurs protéiques qui se recrutent les uns les autres et qui forment avec l'ARNpol II un complexe dit « d'initiation ». Ces facteurs sont appelés TFIIA, TFIIB, ... pour Transcription Factor for RNA polymerase II et correspondent aux facteurs généraux de la transcription puisqu'ils s'assemblent sur tous les promoteurs utilisés par l'ARNpol II.

La liaison du complexe d'initiation au promoteur entraîne l'ouverture et le déroulement des deux brins de son ADN et indique le brin qui va être transcrit.

1.3.1.3 L'élongation

L'ARNpol II est associée à des facteurs protéiques d'élongation. Ces facteurs relâchent la structure de la chromatine et facilitent ainsi la progression de la polymérase. Un ARN pré-messager complémentaire du brin matrice de l'ADN (ou brin antisens), donc identique au brin codant de l'ADN (ou brin sens) aux riboses et uraciles près, commence à être synthétisé selon la direction 5' → 3'.

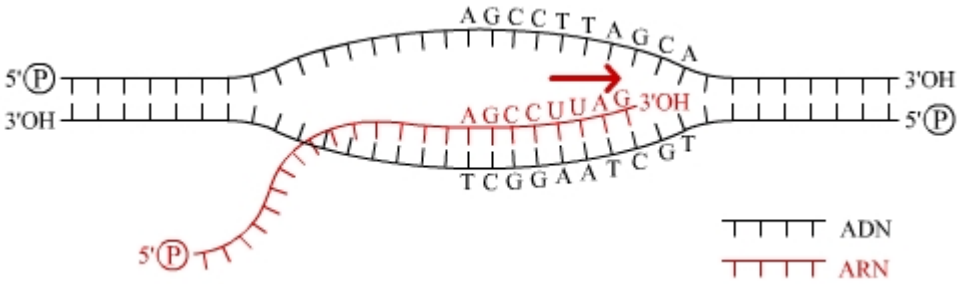


Figure 4 : représentation schématique de l'étape d'élongation de la transcription.

1.3.1.4 La terminaison

L'ARNpol II est équipée de facteurs protéiques de terminaison qui lui permettent de reconnaître un ou plusieurs signaux de terminaison (par exemple TTATTT) portés par le brin parcouru et qui annoncent la fin de la transcription sur le brin d'ADN matrice. La transcription s'arrête alors et l'ARNpol II libère l'ARNpm qu'elle vient d'assembler.

1.3.2 Maturation de l'ARN messager

L'ARNpm subit des modifications avant d'être utilisé pour la synthèse protéique. Toutes ces modifications sont réalisées au fur et à mesure de la synthèse de l'ARNpm. Il existe trois grands types de modifications, catalysées chacune par des enzymes de nature protéique ou ribonucléique :

- ✓ la formation d'une structure particulière en 5' ou coiffe,
- ✓ l'adjonction d'une séquence polyadénylée en 3' ou queue polyA,
- ✓ l'épissage, *i.e.* excision des introns et jonction des exons.

1.3.2.1 L'addition d'une coiffe en 5'

L'addition d'une coiffe a lieu dès le début de la transcription, *i.e.* lorsque la chaîne compte moins de 30 nucléotides. Elle consiste en l'ajout d'une guanine sur l'extrémité 5' de l'ARN, puis de sa méthylation sur l'azote 7 de la base et de la méthylation en 2' du ribose du premier ou des deux premiers nucléotides du transcrit. La particularité de cet ajout consiste dans le type de liaison mis en jeu. Il en résulte que l'extrémité 5' de l'ARNm n'est pas porteuse des trois acides phosphoriques libres habituels, mais d'une guanosine monophosphate ou GMP. Ce phénomène permet de limiter la réactivité de l'extrémité 5' et sa reconnaissance par les exonucléases, protégeant ainsi le transcrit contre la dégradation. Par ailleurs, la coiffe est nécessaire à l'export de l'ARNm vers le cytoplasme et à la liaison de l'ARN avec la petite sous-unité du ribosome lors de l'initiation de la traduction.

1.3.2.2 L'excision et épissage

Les gènes des eucaryotes sont constitués d'une alternance d'exons (parties codantes du gène) et d'introns (parties non codantes, bornées par des séquences de bases spécifiques : 5'GU et 3'AG). Ils sont recopiés intégralement dans l'ARNpm, puis l'ARNpm subit une opération d'excision des introns suivie d'un épissage, *i.e.* la réunion des exons qui constituent l'ARNm. Ce remaniement se déroule au fur et à mesure de la progression de la transcription.

L'épissage n'est pas seulement constitutif ; en effet, il existe des épissages alternatifs, dans lesquels l'élimination des introns (ou de portions d'exons) peut faire se lier entre eux des exons non consécutifs. Ce mécanisme démultiplie les capacités codantes d'un gène.

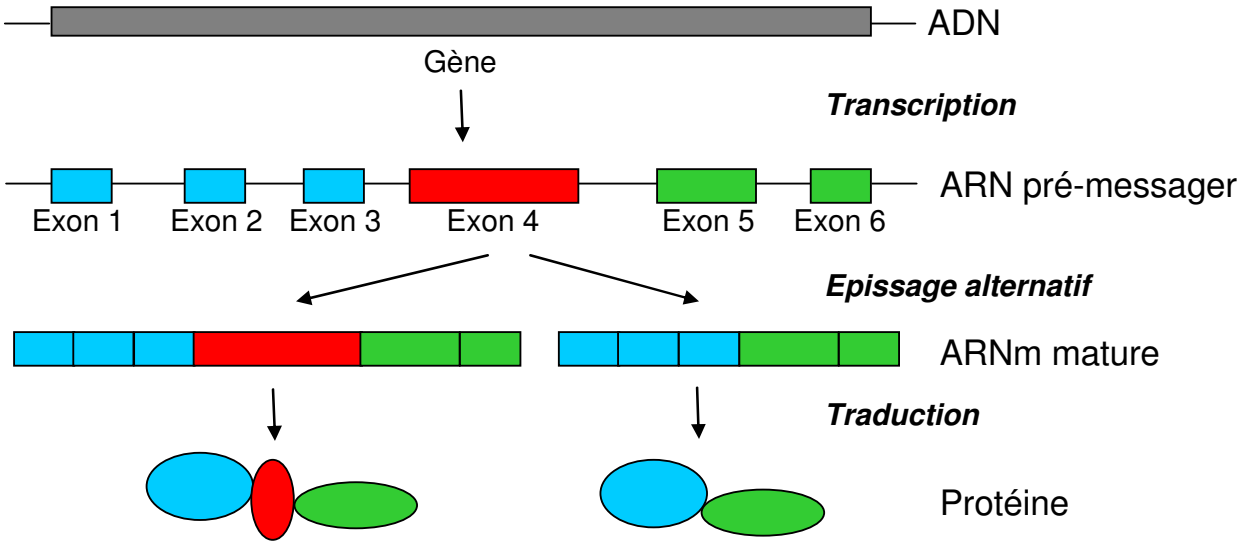


Figure 5 : représentation schématique de l'épissage alternatif.

1.3.2.3 L'addition d'une queue polyA en 3'

Dans les cellules eucaryotes, tous les ARNm possèdent une queue polyA à leur extrémité 3' sauf les ARNm codant les histones. La queue polyA est produite après la transcription par un clivage endonucléolytique au site de polyadénylation, lui-même codé au niveau du gène. Le site de clivage déterminé par le dinucléotide CA est entouré par une séquence AAUAAA très conservée située 10 à 30 nucléotides en amont du site de clivage et par une séquence DSE (DownStream Element), riche en U ou en GU, située environ 30 nucléotides en aval du site de clivage. La séquence AAUAAA est reconnue spécifiquement par le complexe CPSF (Cleavage and Polyadenylation Specific Factor) et la séquence DSE par le complexe CstF (Cleavage Stimulation Factor). Ces deux complexes et d'autres composants, comme l'ARNpol II et la polyA polymérase, interagissent en formant un complexe capable de cliver la molécule d'ARNpm au niveau du site de clivage.

La polyA polymérase ajoute alors environ 200 nucléotides. Notons que cette polymérase n'utilise pas de matrice ADN pour créer cette séquence polyA. La queue polyA n'est par conséquent pas codée par le génome. Elle confère de la stabilité au futur ARNm et est perdue au fur et à mesure qu'il est traduit.

Notons que des protéines spécifiques se lient au niveau de la coiffe (cap-binding protein) et de la queue polyA (polyA-binding protein) de manière à protéger l'ARN de l'action des exoribonucléases.

L'ARNm ainsi généré porte donc une coiffe côté 5' et une queue polyA côté 3'. La séquence d'ARN contient par ailleurs deux parties distinctes, à savoir une partie codante, qui sera traduite en acides aminés, et une partie non codante. La partie non codante ou UTR (untranslated regions) se situe de part et d'autre de la partie codante :

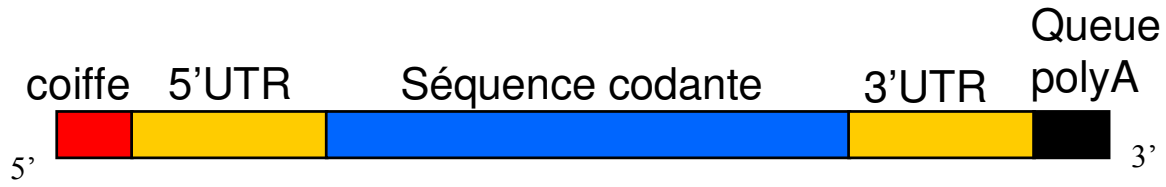


Figure 6 : représentation schématique de l'ARNm mature.

1.3.3 Fidélité de la transcription

L'altération de la séquence nucléotidique d'un ARNm peut résulter de trois mécanismes :

- ✓ le gène codant porte une mutation,
- ✓ une erreur s'est produite lors de la synthèse de l'ARNm ou lors de sa maturation,
- ✓ une modification enzymatique a lieu.

L'altération des séquences ARNm peut avoir un impact codant ou non. Seuls les mécanismes décrits chez l'homme seront présentés. Il existe des mécanismes de surveillance nucléaires et d'autres cytoplasmiques. En effet, un premier niveau de contrôle détecte dans le noyau les défauts de maturation (absence de coiffe, erreurs d'épissage ou encore défauts dans la queue polyA). L'export vers le cytoplasme est alors bloqué. Les ARN reconnus défectueux sont alors dégradés le plus souvent par l'exosome (voie 3' > 5'), complexe de plus de 10 sous-unités [65], ou par une voie 5' > 3' impliquant le clivage de la coiffe. Les ARN ainsi vérifiés sont exportés vers le cytoplasme où ils subissent un autre contrôle appelé NMD (Non-sense mediated decay).

1.3.3.1 Mécanismes de surveillance

- ✓ Non-sense Mediated Decay

Le NMD est un système localisé dans le cytoplasme conduisant à la dégradation et donc à un arrêt prématuré de la traduction des ARN porteurs d'un codon stop situé en amont du stop canonique [66,67]. Les codons stop prématurés (PTC pour Premature Termination Codon)

peuvent être générés par un décalage du cadre de lecture, une mutation non sens, une erreur d'épissage ou une erreur de transcription. Le mécanisme permettant d'identifier les ARN dirigés vers le NMD repose sur les complexes de jonction des exons EJC (Exon Junction Complex) situés en amont des jonctions exon-exon. Les EJC sont assemblés puis déposés par le spliceosome au cours de la maturation de l'ARN [68]. En règle générale, seuls les codons stop situés à une certaine distance de la jonction (plus de 55 nucléotides) sont détectés par le NMD [69]. Le mécanisme du NMD est décrit dans la figure 7 :

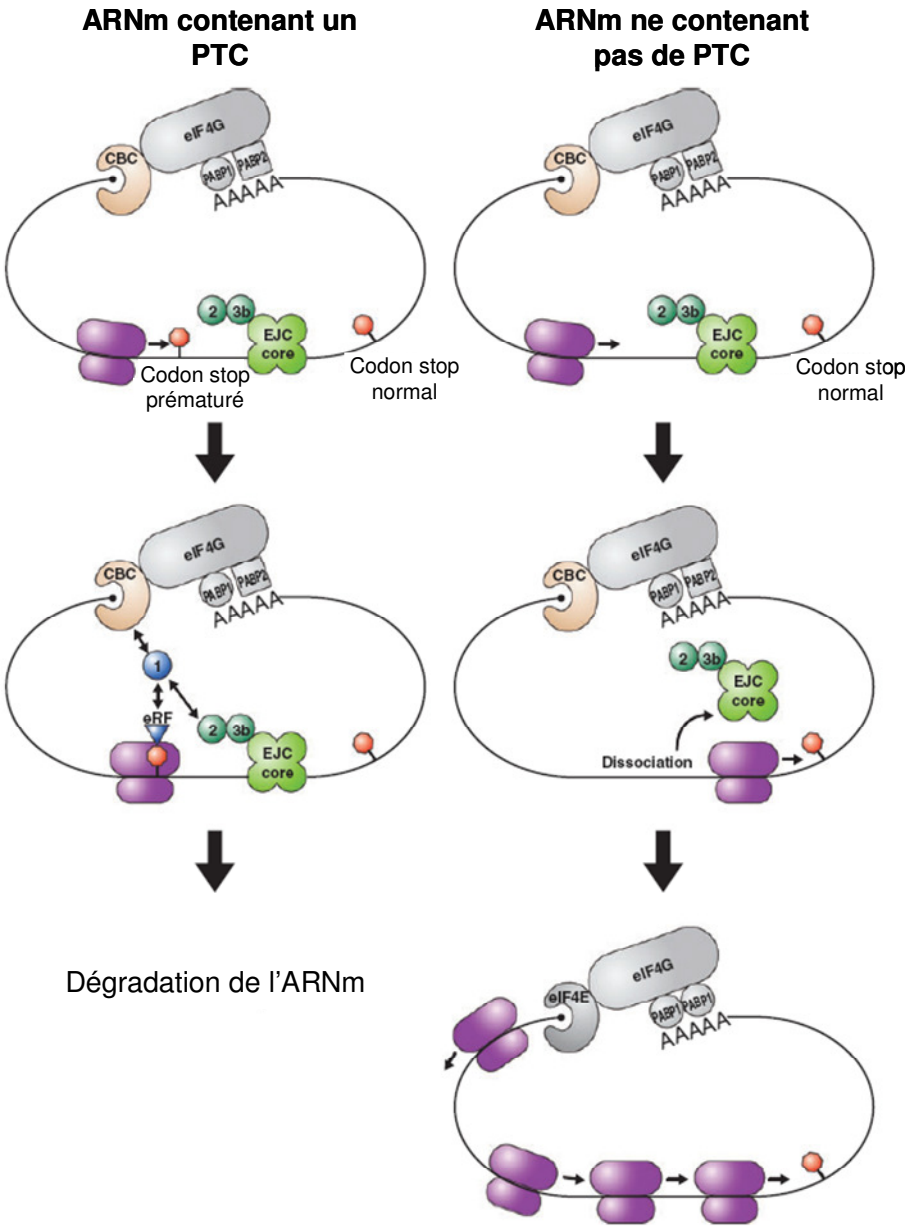


Figure 7 : différentes étapes du Non-sense Mediated Decay chez les mammifères [70].

Un ribosome traduit l'ARNm et s'arrête lorsqu'il rencontre un codon stop. Deux options sont alors envisageables : 1) il s'agit du codon stop canonique (droite) et l'ARN n'est pas dégradé, ce qui implique qu'un autre cycle de traduction peut commencer, 2) le codon stop rencontré est un PTC (gauche) et le NMD intervient. En effet, le NMD est déclenché quand il y a interaction entre la protéine UPF1 (up-frameshift protein 1) associée au ribosome et UPF2, lui-même lié à UPF3b. UPF2 et UPF3b sont associées au complexe EJC. Dans le cas d'un transcrit portant un PTC, il existe au moins un complexe EJC en amont du codon stop naturel, ce qui permet l'interaction entre UPF1, recruté par le complexe CBC (cap-binding complex) et les facteurs eRF1 et eRF3 (eukaryotic release factors), et UPF2 / UPF3b. L'ARNm est donc dégradé. Par opposition, le NMD n'est pas déclenché en présence d'un transcrit ne contenant pas de PTC car les complexes EJC sont situés en amont du codon stop. Les complexes EJC sont dissociés au moment du passage du ribosome.

Il existe également des formes de NMD contrôlant les variants d'épissage appelées NAS (Non-sense-associated altered splicing). Par exemple, il existe des gènes (CAD, IDUA) présentant des codons stop en phase situés dans les introns et induisant, en cas de transcription erronée de cet intron, la dégradation de l'ARN. Cette observation a été étendue à plus de 90 % des gènes étudiés [71].

- ✓ Staufen Mediated Decay (SMD)

Par des mécanismes semblables à ceux du NMD, la protéine Stau1 détecte la présence de ribosomes arrêtés au niveau de PTC et recrute Upf1, facteur impliqué également dans le NMD, et provoque la dégradation de l'ARN lorsque cette liaison est établie dans l'UTR 3', suffisamment en aval du codon stop (≥ 25 nucléotides) [72,73].

Contrairement aux mécanismes du NMD, le facteur Upf1 est recruté directement par Stau1 et non par l'intermédiaire de jonctions EJC.

- ✓ NonStop Decay (NSD)

Le NSD a pour rôle de dégrader les transcrits ne possédant pas de codon stop [74]. L'absence de codon stop peut être due à une polyadénylation prématurée au sein de la partie codante ou à un arrêt prématuré de la transcription. Il est peu probable que le NSD soit lié à une mutation du codon stop naturel dans la mesure où la partie non traduite 3' contient en général un ou plusieurs codons stop.

1.3.3.2 Mécanismes de correction

Thomas *et al.* ont montré en 1998 l'existence d'un mécanisme de correction en cas d'incorporation d'un nucléotide non complémentaire à l'ADN au cours de la transcription par l'ARNpol II humaine [75,76]. Les auteurs montrent que, si l'on force *in vitro* l'incorporation d'une base non complémentaire, celle-ci se fait 500 fois moins vite que l'incorporation de la base correcte. Par ailleurs, en cas de mésappariement, l'incorporation de la base suivante est 15 à 20 fois plus lente. L'activité de clivage de l'ARNpol II agit alors sous la stimulation du polypeptide TFIIS, induisant l'élimination de la base mésappariée et la reprise de la phase d'élongation de la transcription.

Malgré l'existence de différents moyens de surveillance, la transcription infidèle est possible. Le taux d'erreur au cours de la transcription est en effet estimé à une erreur toutes les 1.000 à 10.000 bases [77].

1.3.3.3 Exemples d'infidélité de transcription

- *In vitro*

L'étude *in vitro* de l'ARN polymérase ADN-dépendante du bactériophage T7 montre que les substitutions de base sont les évènements les plus communs [78]. La même équipe montre par ailleurs que, chez le bactériophage T7, la bactérie et la levure, les substitutions survenant lors de la transcription ne sont pas dues à des erreurs d'incorporation de base mais plutôt à des problèmes d'alignement entre le brin d'ADN lu et le brin d'ARN [79]. En effet, le mécanisme proposé par les auteurs est schématisé en figure 8 et peut être décrit de la manière suivante :

- ✓ Transcription fidèle de la base n,
- ✓ Formation d'une boucle impliquant la base n+1 de l'ADN et l'excluant momentanément du brin lu,
- ✓ Transcription fidèle de la base n+2,
- ✓ Réalignement des bases n, n+1 et n+2 impliquant le fait que la base incorporée en face de n+2 se retrouve face à n+1,
- ✓ Transcription fidèle de n+2, …

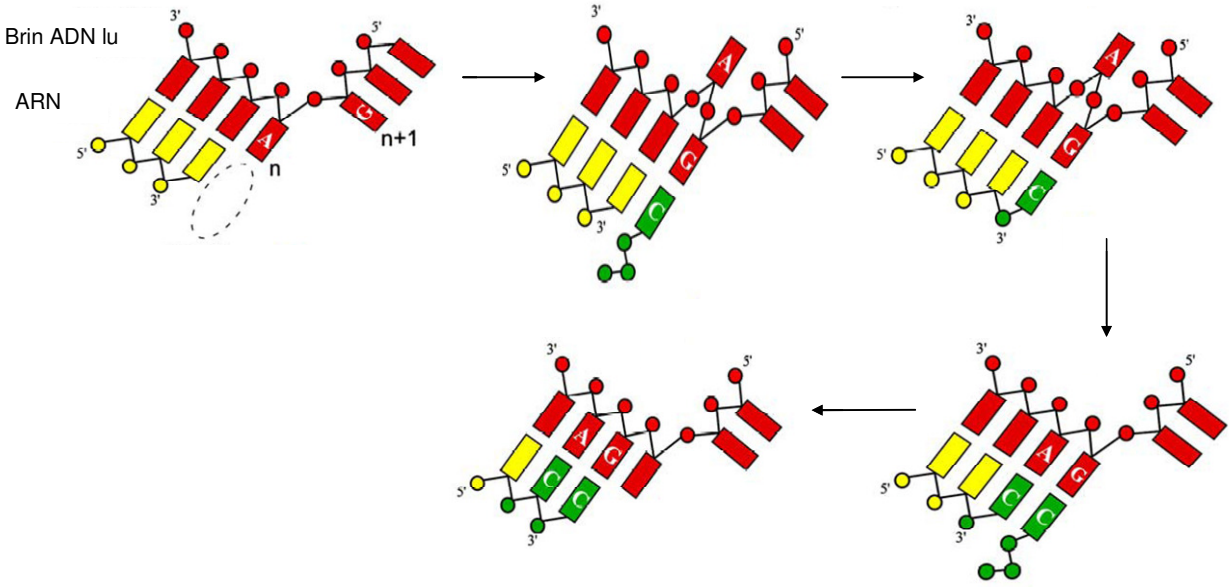


Figure 8 : substitution de base lors de la transcription, d'après Kashkina et al. [79].

- *In vivo*

Quelques cas d'ARN infidèles ont été décrits. Le plus souvent, les cas d'infidélité de transcription ont été découverts lorsque les caractéristiques phénotypiques d'un modèle n'étaient pas en accord avec le génotype responsable de la pathologie.

Dès 1994, l'équipe de Van Leeuwen décrit des délétions GA survenant dans des motifs GAGAG du gène VP (vasopressin) dans des neurones d'un rat souffrant de diabète insipide [80]. Le rat Brattleboro a été choisi puisqu'il porte une mutation germinale induisant la production de VP aberrante. La délétion additionnelle du dinucléotide GA rétablit le cadre de lecture. La délétion du dinucléotide peut survenir à deux positions différentes, d'où la persistance de 13 ou 22 acides aminés décalés. Malgré ces quelques acides aminés divergents de la VP canonique, la délétion GA permet la production de VP fonctionnelle. Il faut noter que le même phénomène est observé chez le rat sauvage, avec toutefois une intensité moindre. Dans ce cas, la délétion GA génère un décalage du cadre de lecture et donc la production de protéines aberrantes, dites protéines +1. L'équipe s'est ensuite intéressée aux gènes humains liés à la maladie d'Alzheimer et a mis en évidence l'existence de délétions GA lors de la transcription des gènes APP (amyloid beta (A_4) precursor protein) et UBB (ubiquitin-B) [81,82]. Ces erreurs de transcription génèrent la production de protéines +1 et sont détectables chez les sujets malades ainsi que chez les sujets témoins de plus de 51 ans uniquement. Ce phénomène

pourrait expliquer pourquoi l'âge est le premier facteur de risque de la maladie d'Alzheimer. L'hypothèse soulevée par les auteurs est la baisse d'efficacité du système NMD avec l'âge.

D'autres études font état de décalages de base lors de la transcription :

- ✓ En 1992, une étude a été réalisée sur des patients atteints d'hypobetalipoprotéinémie due à une délétion de cytosine dans le gène codant l'ApoB (apolipoprotein B) [83]. Cette délétion crée un motif de 8 adénines consécutives et conduit à la synthèse d'une protéine tronquée. La production de protéine normale à partir du gène muté a néanmoins été observée. Les auteurs montrent en effet l'existence d'ARNm (11%) présentant une insertion d'adénine au sein du motif A_8. Cette insertion restaure le cadre de lecture et permet la production d'ApoB canonique.

- ✓ L'étude d'une famille japonaise atteinte d'hémophilie A due à une délétion de thymine au sein du motif A_8TA_2 du gène du facteur VIII a révélé des signes cliniques moins sévères que ceux attendus [84]. En effet, la délétion induit un décalage du cadre de lecture et le gène ne peut donc pas mener à la production de facteur VIII fonctionnel. Or, il y a production, chez ces patients, d'une faible quantité (1 à 7%) de facteur VIII normal. L'analyse des ARN a montré l'existence d'ARN dits mutants. Les ARN mutants présentent 7, 8, 9 ou 11 adénines, là où 10 sont attendues. Les mutants A_8 et A_{11} restaurent le cadre de lecture canonique. A noter que l'événement observé le plus fréquemment est une délétion d'une adénine. Cette délétion est aussi observée chez des individus sains.

- ✓ En 2000, une équipe japonaise a étudié le rat LEC (Long Evans Cinnamon) présentant un métabolisme anormal du cuivre et par conséquent un risque augmenté d'hépatites et cancers du foie [85]. L'étude du gène P53 a montré que 7 à 9% des ARN transcrits à partir de ce gène portaient une insertion d'adénine au sein de régions contenant 6 adénines consécutives. Remarquons que le gène P53 contient un motif constitué de 6 guanines consécutives et qu'aucune insertion n'a été détectée dans ce motif. Notons enfin que le phénomène d'insertion lors de la transcription du gène P53 humain n'a pas été observé ; le gène humain ne contient pas de motif de 6 adénines consécutives.

- ✓ Des cas de neutropénie cyclique ont été étudiés chez le chien. Il s'agit d'une maladie autosomale récessive caractérisée par des variations du nombre de neutrophiles,

variant cycliquement (deux semaines) de zéro à des valeurs normales. La cause de cette pathologie est une insertion d'adénine au sein d'un motif de 9 adénines consécutives du gène AP3B1 (adaptor-related protein complex 3, beta 1 subunit). Benson et al. ont montré, en 2004, que les chiens homozygotes produisaient des ARN mutés et des ARN normaux, contrairement à ce qu'indique leurs génotypes [86]. Ainsi, des erreurs survenant au cours de la transcription provoquent une délétion d'adénine au sein du motif A_{10}, permettant la production d'ARN non porteurs de la mutation, bien que celle-ci soit présente au niveau de l'ADN.

Ainsi, le dogme central de la biologie moléculaire selon lequel la séquence d'ADN permet de prédire la séquence d'ARN a été plusieurs fois ébranlé ces 20 dernières années. Les cas de décalage du cadre de lecture générés au cours de la transcription ont été le plus souvent décrits *in vitro* comme des substitutions et *in vivo* comme des délétions ou insertions survenant au sein de motifs répétés. Les transcrits peuvent donc contenir des erreurs alors que l'ADN ne porte pas de mutation. L'étude des ARN peut donc permettre d'appréhender une forme d'hétérogénéité des cellules. Les données sélectionnées pour étudier *in silico* à grande échelle la fidélité des ARN sont les Expressed Sequence Tags.

1.4 Expressed Sequence Tags

Les ARNm exprimés dans une cellule sont retro-transcrits en ADN complémentaire (ADNc) et séquencés. Les ESTs (Expressed Sequence Tags ou Etiquettes de Séquences Exprimées) sont des fragments d'ADNc [87,88]. Le séquençage d'ADNc consiste donc à caractériser l'ensemble des ARNm exprimés dans une cellule.

1.4.1 Définition et mode d'obtention

La synthèse d'une banque d'ESTs consiste tout d'abord à extraire les ARNm d'un tissu donné. Ensuite, les ADNc, copie des ARNm, sont générés par transcription reverse et amplification à partir de cette population. Puis, les ADNc sont clonés et séquencés partiellement. Le séquençage se fait à partir de l'une ou l'autre des extrémités, voire au sein même de l'ADNc (on parle alors d'amorçage interne [89,90]). Les ESTs sont donc des séquences nucléotidiques de 500 bases environ représentatives du transcriptome exprimé dans un tissu à un moment donné.

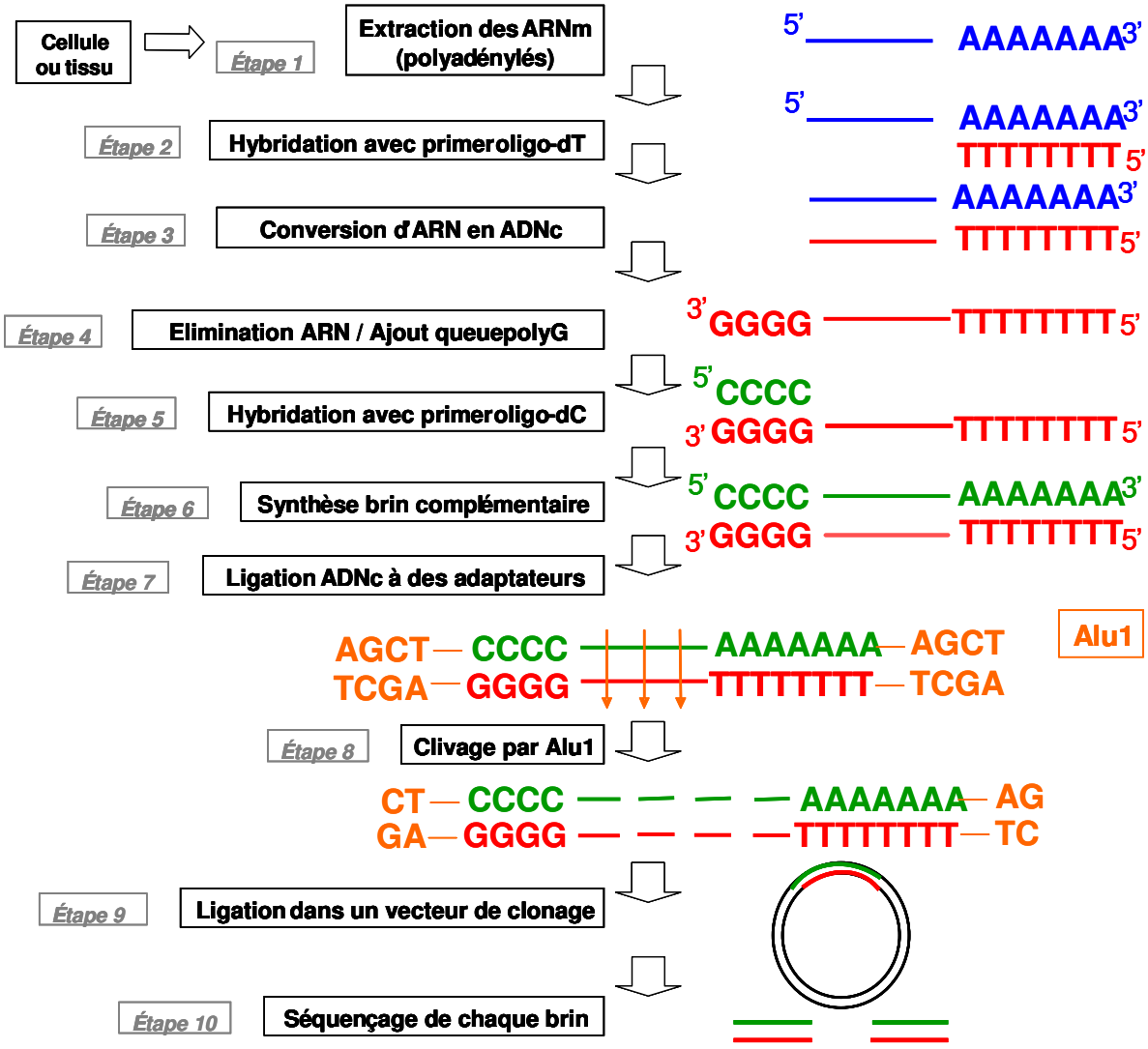


Figure 9 : principe d'obtention des ESTs.

Les ESTs sont ainsi produites à partir de l'ensemble des ARNm présents. Les gènes peu exprimés sont représentés par peu de molécules d'ARNm et donc par moins d'ESTs.

1.4.2 Rôles et limites des ESTs

Le nombre considérable d'ESTs disponibles dans les différentes bases de données en fait une source d'information utile pour :

- La reconstitution, via les contigs, des séquences complètes d'ADNc [87,88,91-96]. Des outils récents, comme ESTAnnotator [97], permettent de contrôler la qualité des ESTs, puis d'identifier et de caractériser de nouveaux gènes de manière automatisée.
- La reconstitution de la structure des gènes [98,99].
- L'identification des variants d'épissage d'un gène [100-102].

- L'étude des régions non traduites, ou UTR, et plus particulièrement des régions promotrices [103,104].
- L'analyse des queues polyA [105-107].
- L'étude de l'expression des gènes [108], voire de l'expression différentielle des gènes entre tissus cancéreux et non cancéreux [109,110].
- L'identification de polymorphismes (Single Nucleotide Polymorphism ou SNP) [111-115].
- La conception des puces à ADN [116,117].

De nombreux outils ont été développés ces dernières années pour faciliter la gestion des ESTs, comme ESTAP [118], PartiGene [119], ESTExplorer [120] ou encore EST2uni [117]. L'un des plus récents, OREST, permet d'affecter une EST au gène ou à la protéine correspondante et permet en outre l'analyse fonctionnelle du jeu de données [121].

Les ESTs présentent aussi des limites :

- Les ESTs sont séquencées en un seul passage, c'est pourquoi elles contiennent des erreurs. En effet, l'objectif initial était la découverte de nouveaux gènes impliqués dans les maladies ; les erreurs de séquençage n'empêchant pas l'identification du gène dont dérive l'EST, elles n'ont pas été supprimées. Cela entraîne des erreurs de séquençage non négligeables (3 à 5%) [89,122]. A noter que le taux d'erreur « standard » pour le projet de séquençage du génome humain est inférieur à 0,01% [123].

  Les régions reconnues de moindre qualité sont plutôt localisées aux extrémités des ESTs, *i.e.* au niveau des 50-100 premières et dernières paires de bases de l'EST [89,124].
- Les éléments assurant l'expression d'un gène tels que les promoteurs et des éléments situés hors des séquences codantes ne peuvent être étudiés.
- Les ESTs ne sont pas strictement représentatives de tous les ARNm du génome puisqu'il existe un biais lié à la technique d'obtention. En effet, les ESTs sont générées par une transcriptase inverse qui synthétise l'ADNc en progressant dans le sens 3' $\rightarrow$ 5' et se décroche souvent avant d'arriver à l'extrémité 5' de l'ARNm. Il y a donc artificiellement un enrichissement des ADNc représentant l'extrémité 3' des ARNm par rapport à ceux représentant leur extrémité 5'.
- Le niveau d'expression des gènes influence le nombre d'ESTs disponibles. Les gènes peu exprimés sont moins bien représentés dans les banques d'ESTs, alors qu'ils peuvent représenter 60% des gènes d'un organisme [108].
- En 1996, Hillier *et al.* se sont intéressés à la qualité des ESTs déposées [125]. Ils ont ainsi montré que les séquences pouvaient être contaminées par de l'ADN bactérien,

mitochondrial, par des introns et des régions inter-géniques, par des chimères combinant deux ADNc ou par du vecteur de clonage.

Pour pallier ce problème, des outils de « nettoyage » ont été développés, comme CleanEST [126], qui consiste à aligner les ESTs à différentes sources de contaminants (vecteurs de clonage UniVec http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html, *Escherichia coli*, ADN mitochondrial).

- En outre, le manque de précision (inévitable en pratique) au moment du prélèvement de certains tissus peut être une source supplémentaire d'erreur. Ainsi, les ESTs issues de tissus cancéreux peuvent être contaminées par des cellules stromales non pathologiques qui forment le tissu de soutien des cellules tumorales.
- Enfin, les ESTs souffrent souvent d'approximation des annotations, voire de manque d'information. Le vocabulaire permettant de décrire les ESTs n'est pas toujours standardisé ; ce manque de standardisation rend l'analyse automatisée complexe.

1.4.3 Banques de données

Malgré les limites préalablement exposées, les ESTs, de part leur utilité, la rapidité et le faible coût d'obtention, ont été générées en quantités très importantes. Pour être facilement accessibles et utilisables, ces séquences ont rapidement été stockées dans des banques de données, publiques et privées.

La principale banque publique est dbEST [127], banque maintenue et distribuée par le NCBI. Aujourd'hui, plus de 55 millions d'ESTs sont disponibles dans dbEST, dont 8,1 millions d'ESTs d'origine humaine. dbEST est une division de Genbank [128], la banque de séquences de référence. Genbank est aussi créée, maintenue et distribuée par le NCBI, et contient l'ensemble des séquences nucléotidiques publiques.

Les entrées dbEST contiennent bien sûr l'information de séquence de l'EST mais aussi des informations concernant la librairie de laquelle est issue l'EST. Cette donnée va s'avérer primordiale dans la mesure où elle permet d'accéder à la nature (cancéreuse ou non) du tissu à partir duquel l'EST est extraite.

dbEST contient des séquences redondantes contenant un nombre non négligeable d'erreurs de séquençage. Ce problème a motivé le développement d'UniGene, outil permettant de regrouper les séquences de bonne qualité et correspondant à un gène [129].

Des clusters de séquences ont été construits pour tous les organismes disposant d'au moins 70.000 ESTs. La première étape de la formation de clusters consiste à éliminer au maximum les séquences de mauvaise qualité ainsi que les séquences issues d'organismes étrangers,

d'ADN mitochondrial ou ribosomal, ou dérivant du vecteur de clonage ou de l'amorce. Les séquences répétitives pouvant conduire à des erreurs d'alignement sont traitées avec précaution. Ainsi, pour qu'une séquence soit incluse dans UniGene, elle doit contenir au moins 100 paires de bases de bonne qualité et non répétitives. Les clusters initiaux sont formés à partir des séquences complètes d'ARNm. Puis, un outil d'alignement compare l'ensemble des ESTs de qualité suffisante aux clusters initiaux. Un cluster correspond donc à l'ensemble des séquences (ARNm et ESTs) disponibles pour un gène, cet ensemble représentant en réalité un sous-ensemble épuré des séquences disponibles pour ce gène. De nombreux outils ont été décrits afin d'assembler les ESTs en clusters, comme Phrap [130] ou CAP3 [131].

D'autres banques, comme stackPACK [132] ou TIGR Gene Indices [133], regroupent les ESTs correspondant au même gène.

Nous avons choisi de considérer l'ensemble des ESTs disponibles (dbEST) et non un sous-ensemble épuré de manière à prendre en compte toute l'hétérogénéité contenue dans ces séquences.

1.5 Formulation de l'hypothèse de travail

Notre premier objectif a été de déterminer si l'analyse des séquences ESTs permet de rendre compte d'une différence d'hétérogénéité entre les cellules tumorales et les cellules normales. En d'autres termes, l'hétérogénéité des ARN cancéreux est-elle plus grande que celle des ARN non cancéreux ?

L'objectif était donc de comparer, transcrit par transcrit, l'hétérogénéité des ESTs issues de tissus normaux et l'hétérogénéité des ESTs issues de tissus cancéreux. L'originalité de cette étude réside dans la comparaison statistique de 2 matrices d'ESTs, à savoir les ESTs dites normales (*i.e.* issues de tissu non cancéreux) et les ESTs dites cancéreuses.

Une étude préliminaire portant sur 17 gènes abondamment exprimés et s'intéressant uniquement aux substitutions de base a permis d'introduire la notion d'infidélité de transcription affectant davantage les tissus tumoraux que les tissus non tumoraux. Cette étude a ensuite été étendue à l'ensemble des transcrits humains, les évènements étudiés étant les substitutions de base, mais aussi les délétions et les insertions. La pertinence statistique des résultats obtenus nous a alors permis de formuler des prédictions bioinformatiques au niveau de l'ARNm et de la protéine, prédictions qui ont été confirmées par les équipes de biologistes.

2 Étude préliminaire

2.1 Démarche bioinformatique

2.1.1 Extraction et tri des ESTs

La banque de données dbEST a été choisie pour réaliser cette étude. L'objectif du travail étant d'analyser l'hétérogénéité des séquences, il nous a paru plus judicieux de considérer un ensemble de séquences non épuré.

Les ESTs d'origine humaine utilisées pour cette étude préliminaire ont été téléchargées le 25 juin 2005 sur le site du NCBI. Les 6.100.538 ESTs ont ensuite été partagées en 3 groupes suivant le tissu à partir duquel elles avaient été extraites. Le système mis en place pour trier les ESTs repose sur l'information contenue dans la description de la librairie à laquelle l'EST appartient et a permis d'obtenir un premier groupe de 2.617.506 ESTs cancéreuses, un deuxième groupe de 2.811.245 ESTs normales et un dernier groupe de 671.787 ESTs dont l'origine n'est pas précisée. Nous avons choisi de ne pas prendre en compte les ESTs dont on ne connaît pas l'origine.

2.1.2 Choix de 17 transcrits

L'objectif de cette étude préliminaire est de savoir si l'information contenue dans les ESTs est capable de rendre compte de l'hétérogénéité inhérente au cancer. Nous avons donc sélectionné, pour cette étude préliminaire, des transcrits abondamment exprimés et donc fortement représentés en termes d'ESTs. Les transcrits choisis sont ceux pour lesquels le nombre d'ESTs définissant le cluster Unigene est le plus important. Aucune information sur la fonction des protéines codées par ces gènes n'a été prise en compte.

Les séquences ARNm de référence (RefSeq [134]) ont été téléchargées depuis le site du NCBI.

ALB	albumin	NM_000477.3
ALDOA	aldolase A, fructose-bisphosphate	NM_000034.2
ATP5A1	ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit 1, cardiac muscle	NM_001001937.1
CALM2	calmodulin 2 (phosphorylase kinase, delta)	NM_001743.3
ENO1	enolase 1, (alpha)	NM_001428.2
FTH1	ferritin, heavy polypeptide 1	NM_002032.2
FTL	ferritin, light polypeptide	NM_000146.3
GAPDH	glyceraldehyde-3-phosphate dehydrogenase	NM_002046.3

HSPA8	heat shock 70kDa protein 8	NM_006597.3
LDHA	lactate dehydrogenase A	NM_005566.1
RPL7A	ribosomal protein L7a	NM_000972.2
RPS4X	ribosomal protein S4, X-linked	NM_001007.3
RPS6	ribosomal protein S6	NM_001010.2
TMSB4X	thymosin beta 4, X-linked	NM_021109.2
TPI1	triosephosphate isomerase 1	NM_000365.4
TPT1	tumor protein, translationally-controlled 1	NM_003295.1
VIM	vimentin	NM_003380.2

Table 2 : Liste des 17 transcrits de l'étude préliminaire et identifiants NCBI.

2.1.3 Alignements

L'outil d'alignement utilisé est le MegaBLAST, dérivant du BLAST. BLAST, acronyme de Basic Local Alignment Search Tool, exploite la méthode de Altschul *et al.* [135] pour rechercher les similitudes entre une séquence fournie et toutes les séquences d'une base de données. MegaBLAST a été choisi car il a été optimisé pour aligner des séquences qui diffèrent peu. MegaBLAST est en outre plus rapide que BLAST.

Les 17 transcrits ont été traités séquentiellement. L'ensemble des ESTs a été aligné successivement à chaque transcrit par l'outil d'alignement MegaBLAST 2.2.13 [136]. Les paramètres de BLAST utilisés sont les paramètres par défaut, excepté :

-q = -2, pénalité d'un mismatch,

-b = 100.000, nombre de sorties,

-p = 90, pourcentage d'identité minimum entre la RefSeq et l'EST,

-W = 16, taille du mot d'encrage,

-F = F, pas de filtre des régions de faible complexité (les queues polyA des ARNm ont donc été supprimées avant d'effectuer l'alignement).

Seuls les alignements de plus de 50 bases ont été retenus.

2.1.4 Analyse des alignements

Les transcrits ont ensuite été analysés position par position tout le long de la séquence ARNm. Pour chaque position sont déterminés le nombre d'ESTs normales et cancéreuses qui se sont alignées, ainsi que le pourcentage des ESTs qui portent une base différente de celle donnée par RefSeq. Ces informations sont obtenues par lecture des alignements ; les programmes correspondant sont codés en Perl. Pour chaque transcrit, un tableau reflétant l'ensemble des informations contenues dans les alignements est construit :

ESTs d'origine cancéreuse					ESTs d'origine non cancéreuse			
A	T	C	G	RefSeq	A	T	C	G
1	0	992	19	C	0	0	709	1
0	0	1013	0	C	0	0	709	0
0	985	27	0	T	0	692	18	0
0	0	1013	0	C	0	0	707	0
0	1017	0	0	T	0	707	0	0
1	1016	1	0	T	0	705	0	0
1	0	1014	0	C	0	0	706	0
0	0	1016	0	C	0	0	704	0
988	2	25	1	A	701	0	3	0

Table 3 : Effectifs d'ESTs normales et cancéreuses portant A, T, C ou G pour un ensemble de positions d'un transcrit donné.

Ce tableau de données brutes permet de construire un tableau contenant les effectifs à chaque position de la séquence de référence :

ESTs d'origine cancéreuse			
Effectif	identité à RefSeq	déviation par rapport à RefSeq	% déviation
1012	992	20	2,0
1013	1013	0	0,0
1012	985	27	2,7
1013	1013	0	0,0
1017	1017	0	0,0
1018	1016	2	0,2
1015	1014	1	0,1
1016	1016	0	0,0
1016	988	28	2,8

ESTs d'origine non cancéreuse			
Effectif	identité à RefSeq	déviation par rapport à RefSeq	% déviation
710	709	1	0,1
709	709	0	0,0
710	692	18	2,5
707	707	0	0,0
707	707	0	0,0
705	705	0	0,0
706	706	0	0,0
704	704	0	0,0
704	701	3	0,4

Table 4 : Déviations par rapport à la séquence de référence pour un ensemble de positions d'un transcrit donné.

Les SNPs sont des variations normales de séquences survenant au niveau de l'ADN. Nous ne devons donc pas les prendre en compte dans notre analyse. Les SNPs sont déposés et répertoriés dans des banques de données : nous avons choisi de nous servir des informations de la banque dbSNP [137] du NCBI et avons téléchargé l'ensemble des SNPs correspondant aux 17 transcrits étudiés. Les positions des ARNm étant identifiées comme des SNPs ne seront pas prises en compte dans la suite de l'analyse.

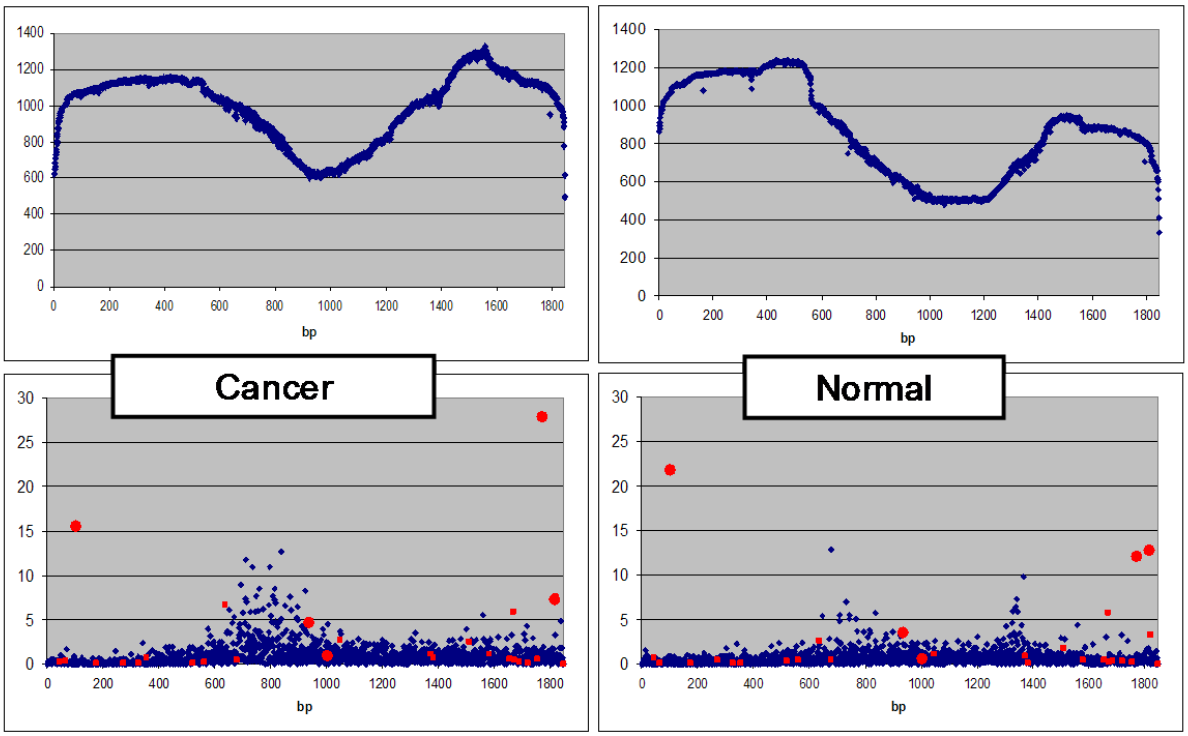


Figure 10 : effectifs d'ESTs cancéreuses (haut gauche) et normales (haut droit) alignées à chaque position de la séquence de référence de VIM, et pourcentage d'ESTs déviant à chaque position (bas).

Les SNPs putatifs sont représentés par des carrés rouges et les SNPs validés biologiquement par des cercles rouges.

L'analyse du transcrit du gène de la vimentine (VIM) montre que chaque position est représentée par au moins 600 ESTs cancéreuses et au moins 500 ESTs normales. La couverture du transcrit en ESTs est donc satisfaisante. La déviation de base par rapport à la séquence de référence présente un bruit de fond de l'ordre de 1 à 3%. Certaines régions du transcrit sont plus variables, des positions pouvant présenter jusqu'à 15% des ESTs alignées portant une base différente de celle de la référence. Ces évènements, que nous nommerons substitutions, sont concentrés dans des régions spécifiques du transcrit, différentes selon l'origine normale ou tumorale des ESTs. Il est important de noter que les plus fortes déviations observées, tant sur les ESTs normales que cancéreuses, sont des positions référencées comme étant des SNPs validés biologiquement.

L'ensemble des graphiques représentant les effectifs et pourcentages de déviation des 17 transcrits sont présentés en annexe 1.

La couverture en ESTs est très satisfaisante et une simple observation des graphiques de déviation suggère que les ESTs cancéreuses portent plus de déviations que les ESTs normales.

2.2 Démarche statistique

2.2.1 Choix du test statistique

L'objectif de cette analyse est de comparer, pour chaque position de chaque transcrit, les déviations portées par les ESTs normales à celles portées par les ESTs cancéreuses. Les ESTs étant obtenues par un séquençage en un seul passage, elles contiennent des erreurs de séquençage. Nous faisons l'hypothèse que les erreurs de séquençage surviennent avec la même intensité dans le séquençage d'une EST normale que d'une EST tumorale. En comparant, à une position donnée, les déviations des deux blocs d'ESTs, il est donc possible de s'affranchir du problème des erreurs de séquençage. L'originalité du travail repose donc non pas sur la détection de substitutions mais sur la comparaison des proportions de substitutions entre les ESTs normales et les ESTs cancéreuses. Un test de proportions est donc effectué pour chaque position de chaque transcrit.

Le test utilisé est le test de comparaison de probabilités unilatéral suivant :

$$\begin{cases} H_0 : p_{1B} \geq p_{2B} \\ H_1 : p_{1B} < p_{2B} \end{cases}$$

où p_{1B} (resp. p_{2B}) représente la probabilité pour une EST cancéreuse (resp. normale) d'avoir la base B à cette position de la séquence de référence. Pour chaque position, le tableau de contingence est construit de la manière suivante :

	B	$\overline{B}$	Somme
Cancéreux	$n_1 - k_1$	k_1	n_1
Sains	$n_2 - k_2$	k_2	n_2
Somme	m_1	m_2	n

où k_1 (resp. k_2) est le nombre d'ESTs ne portant pas la base B parmi les n1 (resp. n2) ESTs cancéreuses (resp. normales) alignées.

La statistique de test T est la suivante :

$$T = \frac{\frac{K_1}{n_1} - \frac{K_2}{n_2}}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

où

$$\hat{P} = \frac{K_1 + K_2}{n_1 + n_2}$$

Sous les conditions de validité :

$$n > 70,\ \frac{n_i m_j}{n} > 5,\ i = 1, 2,\ j = 1, 2,$$

asymptotiquement sous H0, T suit une loi N(0,1).

La règle de décision est définie de la manière suivante :

Si la p-valeur

$$p = P(T > t \text{ sous } H_0)$$

est inférieure à α, alors l'hypothèse H0 est rejetée au seuil α ; sinon l'hypothèse H0 n'est pas rejetée au seuil α.

Si le test est positif (*i.e.* H0 est rejetée), on conclut à une variabilité plus forte des ESTs provenant de tissus cancéreux que des tissus sains (position dite C>N). Par analogie, on définit les positions N>C. Les positions C>N et N>C sont donc déduites des 2 tests unilatéraux.

Une position est dite CndN (ou C non différent de N) si les contraintes du test de proportions sont respectées et que le test bilatéral est non significatif. L'hétérogénéité est alors présente dans des proportions comparables au sein des ESTs normales et cancéreuses, mais on ne peut pas, dans ce cas, exclure l'implication des erreurs de séquençage.

La représentation de la statistique du test permet de définir plus clairement les 3 classes de position :

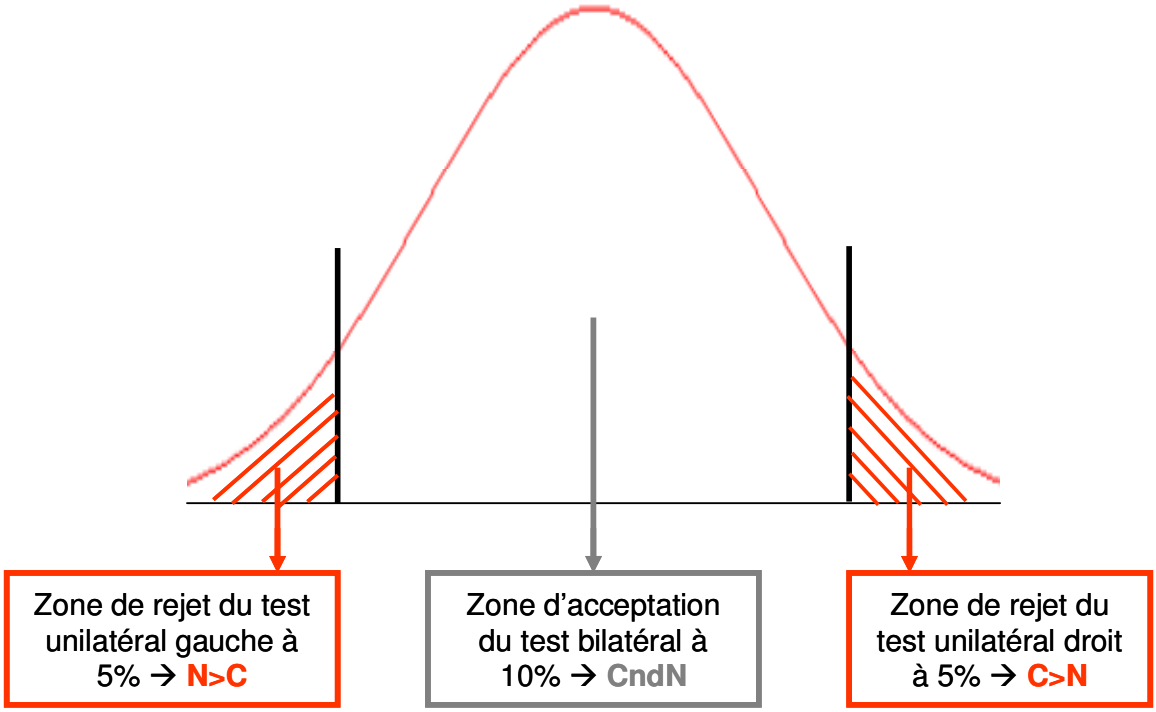


Figure 11 : représentation schématique de la statistique du test de proportions.

En reprenant l'exemple présenté dans le paragraphe 2.1.4, on obtient :

ESTs d'origine cancéreuse			ESTs d'origine non cancéreuse			**test Chi²**
Effectif	identité à RefSeq	déviation par rapport à RefSeq	Effectif	identité à RefSeq	déviation par rapport à RefSeq	
1012	992	20	710	709	1	**C>N**
1013	1013	0	709	709	0	ND
1012	985	27	710	692	18	**CndN**
1013	1013	0	707	707	0	ND
1017	1017	0	707	707	0	ND
1018	1016	2	705	705	0	ND
1015	1014	1	706	706	0	ND
1016	1016	0	704	704	0	ND
1016	988	28	704	701	3	**C>N**

Table 5 : déviations par rapport à la séquence de référence pour quelques positions du transcrit VIM.

ND signifie que le test n'est pas réalisable car les contraintes ne sont pas respectées.

2.2.2 Détermination du nombre de faux positifs

Un test statistique est réalisé pour toutes les positions respectant les contraintes du test de proportions, soit 7.648 tests pour l'ensemble des 23.930 positions des 17 transcrits étudiés.

Les tests unilatéraux sont réalisés au seuil α de 5%, seuil qui correspond au risque d'accepter H1 alors que H0 est vraie. Il est donc indispensable d'estimer le nombre moyen de faux positifs attendus.

Pour une séquence de référence de m positions, on réalise m tests au seuil α. On dispose donc d'autant de p-valeurs p_k, $1 \leq k \leq m$. Le LBE (Location Based Estimator) est une surestimation du nombre moyen de faux-positifs [138] :

$$LBE = 2\alpha \sum_{k=1}^{m} p_k$$

Les tests statistiques sont réalisés pour chacun des 17 transcrits et permettent de caractériser l'hétérogénéité de ces transcrits dans un contexte cancéreux ou non cancéreux.

2.3 Différence d'hétérogénéité des ESTs issues de tissus cancéreux ou normaux

Pour chaque transcrit sont déterminés le nombre de positions pour lesquelles les contraintes du test de proportions sont respectées, le nombre de positions statistiquement significatives C>N et la surestimation du nombre de faux positifs associée et enfin le nombre de positions statistiquement significatives N>C et la surestimation correspondante du nombre de faux positifs :

symbole NCBI	nombre de tests effectués		nombre de tests C>N	LBE		nombre des tests N>C	LBE
ENO1	614		186	18		31	42
FTH1	592		226	20		57	38
GAPDH	812		311	23		61	57
HSPA8	513		163	13		18	37
RPL7A	337		103	11		33	22
RPS4X	371		124	11		27	25
RPS6	432		86	17		40	25
TPT1	489		145	15		26	33
VIM	752		269	24		78	50
ALB	194		38	10		56	8
FTL	678		118	31		86	36

TMSB4X	352		70	18		74	17
ALDOA	336		81	11		35	21
ATP5A1	238		123	3		6	20
CALM2	374		69	16		40	21
LDHA	185		70	4		13	13
TPI1	379		99	14		47	23

Table 6 : résultats des tests statistiques des 17 transcrits étudiés.

Pour l'ensemble des transcrits étudiés sont obtenues 2281 positions C>N (LBE = 259) et 728 positions N>C (LBE = 488). Il existe donc 3,1 fois plus de positions pour lesquelles l'hétérogénéité est plus importante dans les ESTs cancéreuses que normales et ce rapport est de 8,4 si l'on retranche l'estimation du nombre de faux positifs.

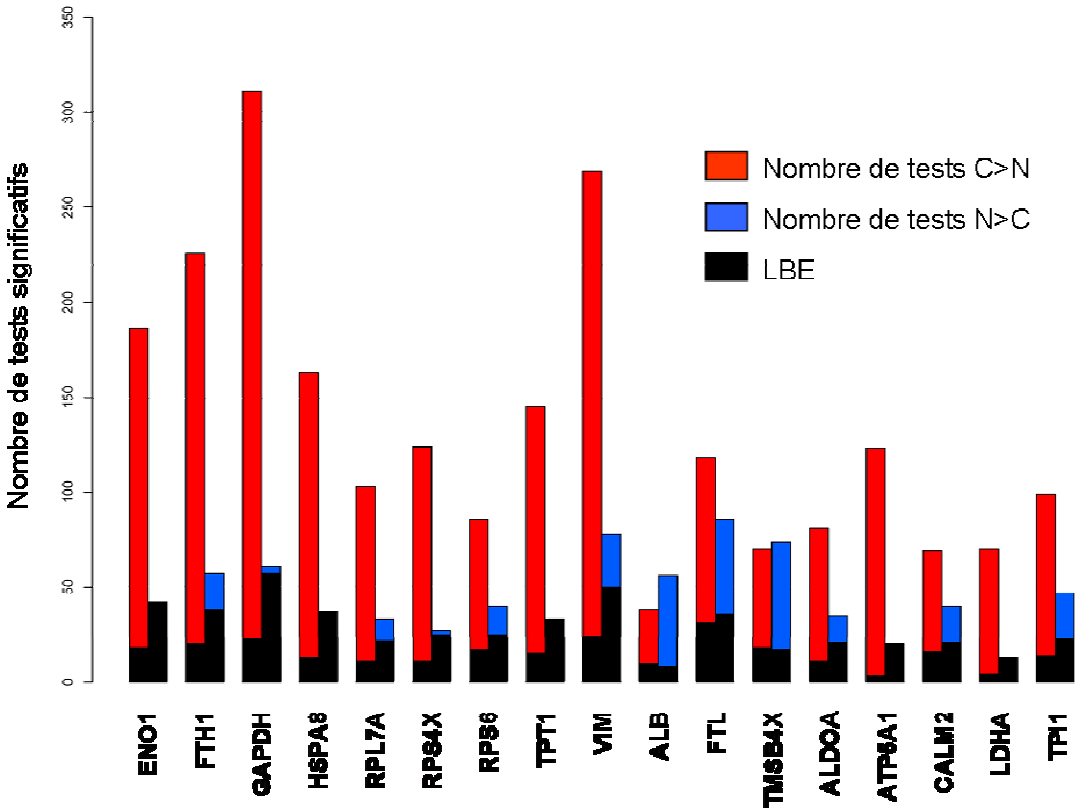


Figure 12 : représentation des résultats statistiques de l'étude préliminaire.

Nombre de tests C>N (rouge) et N>C (bleu) pour chaque transcrit étudié. La part du LBE est représentée en noir.

Pour chacun des 17 transcrits, le nombre de tests C>N excède largement le nombre de faux positifs attendus. Le nombre de tests N>C au contraire est du même ordre que le LBE, voire

inférieur. Pour 13 transcrits, le nombre de positions C>N est plus de deux fois plus grand que le nombre de tests N>C.

Il est donc possible de conclure que le nombre de positions pour lesquelles les ESTs cancéreuses présentent significativement plus de substitutions que les ESTs normales est :

- largement supérieur au bruit de fond,
- supérieur au nombre de positions pour lesquelles la variabilité des ESTs normales est la plus grande (pour 15 des 17 transcrits étudiés).

Nous nous sommes ensuite intéressés à l'importance de la nature des bases situées en amont et en aval des évènements C>N mis à jour.

2.4 Étude du contexte d'ADN

Les substitutions C>N sont concentrées pour la plupart des transcrits dans des régions de 400 à 500 nucléotides, d'où l'idée d'étudier le contexte d'ADN flanquant les substitutions. Pour cela, chaque position de substitution C>N est définie comme l'origine b0 et l'on extrait les 20 bases précédant (b-1 à b-20) et les 20 suivant (b+1 à b+20) la substitution au niveau de l'ARN pré-messager. 2281 séquences de 21 nucléotides dites hétérogènes sont ainsi extraites.

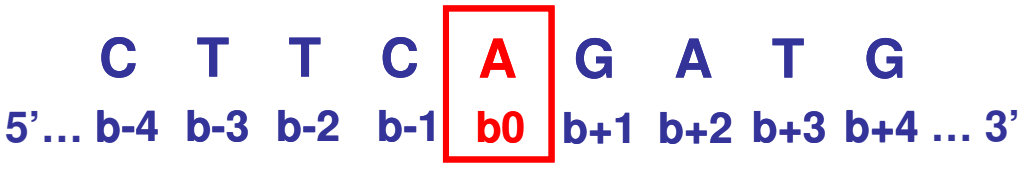
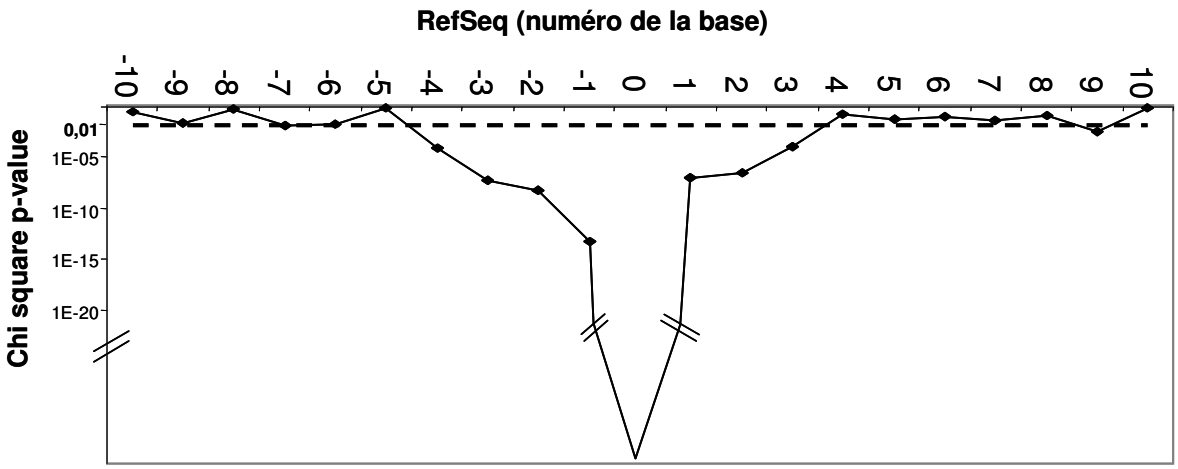


Figure 13 : contexte entourant la base b0 affectée par une substitution.

Les positions non hétérogènes sont définies comme étant des positions pour lesquelles la déviation dans les ESTs cancéreuses est inférieure à 0,5% et pour lesquelles les proportions de substitutions ne sont pas statistiquement différentes entre les deux matrices d'ESTs. 12.273 séquences dites non hétérogènes sont extraites.

Un test de comparaison de proportions de la composition en bases A, T, C, G est réalisé à chacune des 21 positions décrites entre l'ensemble hétérogène et l'ensemble non hétérogène. Les p-valeurs des tests sont représentées ci-dessous.



	b-4		b-3		b-2		b-1		b0		b+1		b+2		b+3	
	H%	NH%	H%	NH%	H%	NH%	H%	NH%	H%	NH%	H%	NH%	H%	NH%	H%	NH%
A	28	24	24	27	25	27	30	22	33	23	29	25	27	26	28	24
T	23	25	28	23	26	23	22	27	32	22	22	24	24	26	24	24
C	24	25	25	24	21	26	23	26	21	26	21	25	20	24	22	26
G	25	26	23	26	28	24	25	25	14	29	28	26	29	24	26	26

Figure 14 : Influence du contexte d'ADN autour de b0.

Effet du contexte de bases sur les événements de substitution. Résultats du test du Chi² pour la position b0 et les positions voisines entre la composition des événements C>N (n=2281 H) et la composition des événements non hétérogènes (n=12273 NH). La ligne en pointillé représente le seuil de significativité (p=0.01). Le tableau indique les compositions de bases significativement différentes entre l'ensemble des sites hétérogènes et l'ensemble des sites non hétérogènes.

La composition des bases flanquant les substitutions C>N est statistiquement singulière. Les bases b-10 à b-5 et b+4 à b+10 ne présentent pas de différence de composition entre sites hétérogènes et sites non hétérogènes, au contraire des bases b-4 à b+3.

Ainsi, les 4 bases ne sont pas substituées dans des proportions uniformes : b0 = A (33%) ~ T (32%) >> C (21%) >> G (14%). De plus, la composition des 4 bases précédant et des 3 bases suivant les événements de substitution est significativement différente de la composition environnant un ensemble de bases non substituées.

La survenue d'un événement d'infidélité de transcription est donc déterminée par la nature de la base subissant la substitution ainsi que par la nature des bases voisines. Il est particulièrement intéressant de noter que les 4 bases précédant et les 3 suivant la base en cours de transcription correspondent au fragment d'ADN ouvert lors du passage de l'ARNpol II [139].

L'infidélité de transcription ne survient donc pas de manière aléatoire, mais est dépendante du contexte d'ADN. En outre, la nature de la base affectée n'est pas répartie de manière uniforme et, comme nous allons le montrer, la nature de la base de remplacement n'est pas non plus aléatoire.

2.5 Règles de remplacement

Une substitution est définie comme le remplacement de la base codée par l'ADN par une autre base. Une substitution C>N affecte une plus grande proportion d'ESTs cancéreuses que d'ESTs normales. On appelle base de remplacement la nature de la base qui est portée par le plus grand nombre d'ESTs cancéreuses déviantes. Pour chaque substitution C>N, seule la base de remplacement majoritaire est prise en considération.

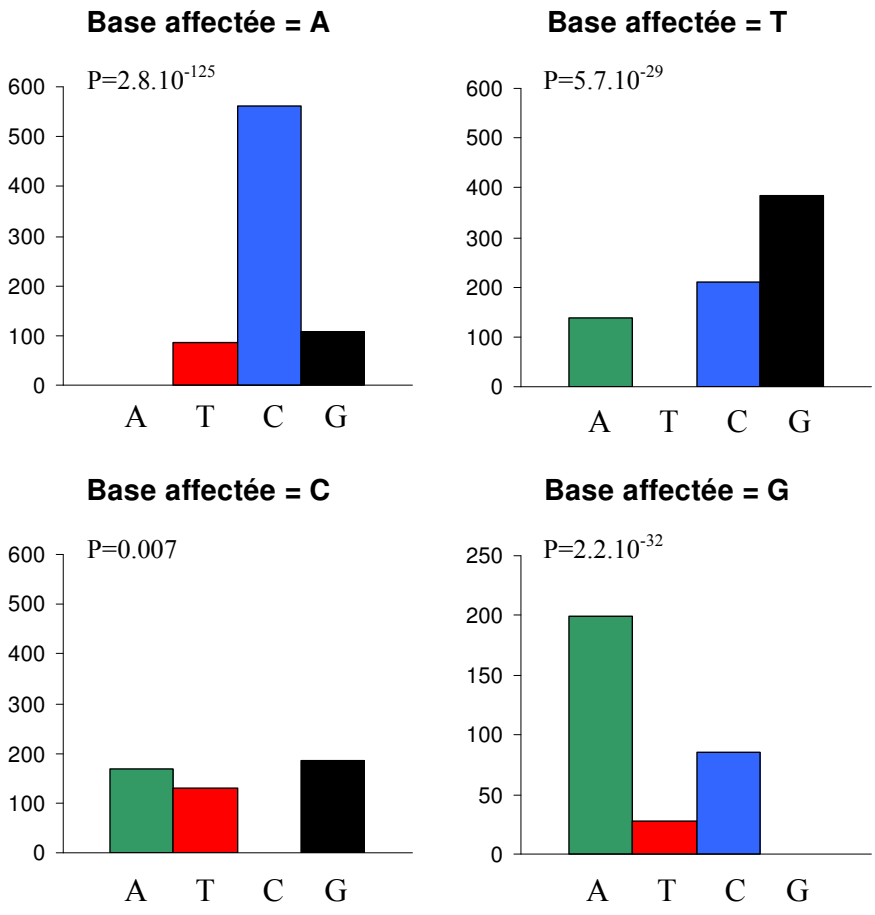


Figure 15 : Nature de la base de remplacement en fonction de la nature de la base affectée.

Les évènements de substitution surviennent dans un contexte non aléatoire (cf paragraphe précédant) ; de plus, chaque base n'est pas substituée uniformément : A est préférentiellement

remplacé par C, T par G, G par A et, dans une moindre mesure, C est rarement remplacé par T. Il existe donc des bases de remplacement préférentielles selon la nature de la base affectée. Cette observation est statistiquement significative ; les p-valeurs données sur la figure 15 correspondent au test d'ajustement des bases de remplacement par rapport à la distribution uniforme.

La base de remplacement n'étant pas sélectionnée de manière uniforme, nous avons ensuite cherché à savoir si sa nature était dépendante du contexte d'ADN (b-1 et b+1). Pour cela, nous avons défini 2 types de situation :

- ✓ Événements informatifs, *i.e.* b0 ≠ b-1 et/ou b0 ≠ b+1 :
  - La base de remplacement Br est de même nature que la base précédant la substitution (c'est-à-dire Br = b-1 et b0 = b+1 ≠ b-1),
  - La base de remplacement Br est de même nature que la base suivant la substitution (c'est-à-dire Br = b+1 et b0 = b-1 ≠ b+1),
  - La base de remplacement Br est de même nature que la base précédant et suivant la substitution (c'est-à-dire Br = b+1, b0 ≠ b-1, b0 ≠ b+1 et b-1 = b+1),
  - La base de remplacement Br est différente de la base précédant et suivant la substitution (c'est-à-dire Br ≠ b+1, Br ≠ b-1, b0 ≠ b-1 et/ou b0 ≠ b+1).

- ✓ Évènements non informatifs, *i.e.* b-1 = b0 = b+1. Dans ce cas bien sûr, la base de remplacement ne peut pas être la même que la base précédant ou suivant la substitution.

Toutes les déviations (n=2281)

base affectée \ base remplacement	A	T	C	G	%
A		85	560	107	33
T	139		210	384	32
C	169	130		185	21
G	199	28	85		14
%	22	11	37	30	

Déviations vers b-1 (n=799)

base affectée \ base remplacement	A	T	C	G	%
A		41	201	42	36
T	28		56	83	21
C	77	41		126	30
G	75	9	20		13
%	23	11	35	31	

Déviations vers b+1 (n=530)

base affectée \ base remplacement	A	T	C	G	%
A		24	52	37	21
T	27		45	177	47
C	38	48		15	19
G	40	11	16		13
%	20	16	21	43	

Déviations ni vers b-1 ni vers b+1 (n=347)

base affectée \ base remplacement	A	T	C	G	%
A		7	148	8	47
T	42		50	9	29
C	17	10		23	14
G	7	2	24		10
%	19	5	64	12	

Déviations vers b-1 = b+1 (n=339)

base affectée \ base remplacement	A	T	C	G	%
A		5	52	12	20
T	11		12	95	35
C	25	30		17	21
G	71	1	8		24
%	31	11	21	37	

Déviations dans un triplet (n=266)

base affectée \ base remplacement	A	T	C	G	%
A		8	107	8	46
T	31		47	20	37
C	12	1		4	6
G	6	5	17		11
%	19	5	64	12	

Figure 16 : Nature des bases affectées et des bases de remplacement en fonction du contexte jouxtant la base affectée.

Ces données montrent que :

- Lors d'un remplacement par la base précédente b-1, A est le plus souvent affecté et remplacé par C. Ainsi CA → CC.
- Lors d'un remplacement par la base suivante b+1, T est le plus souvent affecté et remplacé par G. Ainsi TG → GG. Le schéma est identique lorsque l'on considère le cas où b-1 ≠ b0 et b-1 = b+1.

Ces deux cas permettent de montrer que la base de remplacement est le plus souvent de même nature que la base précédant ou suivant la base affectée.

- Lors d'un remplacement par une base autre que celle précédant et/ou suivant la substitution, alors A est le plus souvent affecté et remplacé par C.
- Lors d'une substitution survenant dans un triplet (b-1 = b0 = b+1), alors A est le plus souvent affecté et remplacé par C.

  Ainsi AAA → ACA.

Ainsi, au cours d'évènements d'infidélité de transcription, la nature des bases affectées et des bases de remplacement n'est pas aléatoire.

Ces conclusions sont néanmoins rendues critiquables de part la nature même (séquençage *single run*) des données exploitées, d'où la mise en place de filtres drastiques des données et alignements générés pour s'assurer de la réalité du signal observé.

2.6 Application de filtres bioinformatiques

Comme cela a été décrit dans l'introduction (paragraphe 1.4.2), le mode d'obtention des ESTs confère à ces données différents inconvénients, comme les erreurs de séquençage, la présence d'ESTs chimériques ou encore la variabilité de longueur des séquences obtenues. Il est donc important de mettre en place différents filtres permettant de tenir compte de ces limites.
Les filtres appliqués consistent d'une part à éliminer de l'analyse les ESTs s'alignant aux transcrits d'intérêt mais ne correspondant pas à ces transcrits et d'autre part à éliminer les positions des transcrits susceptibles de contenir de l'information biaisée. Les filtres sont appliqués séquentiellement.

2.6.1 Filtre des ESTs chimériques, des homologues et des pseudogènes

Les ESTs chimériques dérivent de plusieurs ARN différents liés par erreur. Nous avons filtré les alignements de manière à ne conserver que les ESTs alignées en une seule fois, et ce sur au moins 70% de leur longueur (F1). Ce filtre permet d'éliminer les ESTs chimériques et une partie des ESTs correspondant à des séquences homologues ou à des pseudogènes. Afin d'être le plus stringent possible, nous avons recherché, pour chacun des 17 transcrits, l'ensemble des séquences ARN des gènes paralogues et des pseudogènes.
Les pseudogènes sont identifiés à partir du site *pseudogene.org* et les séquences paralogues sont obtenues par alignement de l'ensemble des transcrits RefSeq (BLASTN, MegaBLAST, Discontigous MegaBLAST). Les ESTs présentant un alignement meilleur contre un paralogue ou un pseudogène que contre le transcrit étudié sont éliminées.
Les tests statistiques sont réalisés sur les données filtrées. Les résultats sont récapitulés à la fin de la partie 2.6 (F2). Ce filtre n'a pas d'impact sur la proportion du nombre de tests C>N par rapport au nombre de tests N>C (rapport C/N).

2.6.2 Filtre des extrémités des alignements

Les 50 premiers et derniers éléments de chaque alignement ont été supprimés (F3). Ce filtre se justifie par le fait que les débuts et fins de séquençage sont de moindre qualité [89,124].

L'application de cette contrainte supplémentaire provoque une baisse importante du nombre de tests significatifs mais une augmentation du rapport C/N.

2.6.3 Normalisation des longueurs des ESTs

Après application des trois premiers filtres, il apparaît que la longueur moyenne des alignements d'ESTs issues de tissus cancéreux est plus grande que la longueur moyenne des ESTs issues de tissus normaux (640 ± 248 et 554 ± 229 bases, respectivement). Nous avons donc normalisé les longueurs des alignements pour chaque transcrit (F4) en tronquant les ESTs d'origine cancéreuse dont la longueur était supérieure à la longueur moyenne d'une EST d'origine normale.

Ce filtre est très stringent et contestable puisque l'on se prive volontairement d'une part importante de l'information située dans la partie la plus hétérogène des transcrits. Le rapport C/N reste néanmoins supérieur à 2.

	F1				F2				F3				F4			
	C>N	**LBE**	**N>C**	**LBE**	**C>N**	**LBE**	**N>C**	**LBE**	**C>N**	**LBE**	**N>C**	**LBE**	**C>N**	**LBE**	**N>C**	**LBE**
ENO1	**186**	18	**31**	42	**165**	17	**31**	38	**94**	8	**17**	21	**52**	7	**21**	14
FTH1	**226**	20	**57**	38	**205**	19	**72**	33	**158**	12	**48**	26	**100**	13	**52**	20
GAPDH	**311**	23	**61**	57	**297**	22	**67**	52	**207**	14	**25**	39	**65**	22	**90**	21
HSPA8	**163**	13	**18**	37	**139**	11	**23**	30	**66**	5	**11**	14	**66**	5	**11**	14
RPL7A	**103**	11	**33**	22	**94**	9	**34**	18	**57**	6	**21**	11	**47**	5	**18**	10
RPS4X	**124**	11	**27**	25	**112**	10	**23**	22	**71**	5	**13**	13	**64**	5	**12**	12
RPS6	**86**	17	**40**	25	**76**	15	**43**	23	**42**	10	**26**	13	**32**	10	**26**	12
TPT1	**145**	15	**26**	33	**137**	14	**28**	30	**78**	9	**14**	20	**78**	9	**14**	20
VIM	**269**	24	**78**	50	**232**	17	**49**	42	**148**	8	**27**	25	**148**	8	**27**	25
ALB	**38**	10	**56**	8	**24**	6	**45**	4	**10**	3	**23**	1	**10**	3	**23**	1
FTL	**118**	31	**86**	36	**121**	29	**91**	34	**77**	22	**56**	24	**57**	23	**70**	20
TMSB4X	**70**	18	**74**	17	**56**	15	**68**	14	**34**	8	**32**	8	**34**	8	**32**	8
ALDOA	**81**	11	**35**	21	**73**	10	**29**	19	**50**	5	**11**	11	**50**	5	**11**	11
ATP5A1	**123**	3	**6**	20	**111**	2	**3**	18	**63**	1	**0**	9	**29**	1	**0**	5
CALM2	**69**	16	**40**	21	**67**	13	**36**	19	**53**	9	**29**	14	**53**	9	**29**	14
LDHA	**70**	4	**13**	13	**60**	3	**8**	12	**31**	1	**2**	5	**31**	1	**2**	5
TPI1	**99**	14	**47**	23	**137**	14	**28**	30	**62**	6	**21**	13	**23**	6	**24**	8
Somme	**2281**	259	**728**	488	**2106**	226	**678**	438	**1301**	132	**376**	267	**939**	140	**462**	220

Table 7 : Nombre de tests significatifs et estimation du nombre de faux positifs associés pour chacun des 17 transcrits après application successive des différents filtres.

	C>N	LBE	N>C	LBE
F1	**2281**	259	**728**	488
F2	**2106**	226	**678**	438
F3	**1301**	132	**376**	267
F4	**939**	140	**462**	220

(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) / LBE]
3,1	8,4
3,1	7,8
3,5	10,7
2,0	3,3

Table 8 : Évolution du rapport entre les tests C>N et les tests N>C de l'ensemble des transcrits après application successive des filtres.

Notons tout de même le danger potentiel de l'excès de filtres. Il faut bien sûr vérifier la pertinence du signal observé, mais si nous filtrons toutes les positions déviant de la référence, alors nous ferons disparaître le signal.

2.6.4 ESTs issues de lignées

L'ensemble des ESTs d'origine cancéreuse provenant de cultures cellulaires ont été éliminées de l'analyse. Après application de cette contrainte, on dénombre 1.009 positions C>N (LBE = 117) et 445 positions N>C (LBE = 193). Le rapport C/N reste donc supérieur à 2 : le signal mesuré n'est pas dû aux ESTs issues de lignées cancéreuses.

PUBLICATION 1 :

Nonrandom variations in human cancer ESTs indicate that mRNA heterogeneity increases during carcinogenesis.

Brulliard M, Lorphelin D, Collignon O, Lorphelin W, Thouvenot B, Gothié E, Jacquenet S, Ogier V, Roitel O, Monnez JM, Vallois P, Yen FT, Poch O, Guenneugues M, Karcher G, Oudet P, Bihain BE.

Proc Natl Acad Sci U S A. 2007 May 1;104(18):7522-7. Epub 2007 Apr 23.

L'étude préliminaire a permis l'étude approfondie de 17 transcrits abondamment exprimés et la formulation de l'hypothèse d'infidélité de transcription, décrite comme un phénomène augmenté dans les tissus cancéreux par rapport aux tissus normaux et dont le contexte d'ADN n'est pas aléatoire. Cette première étude s'est intéressée uniquement aux substitutions de base. Par ailleurs, les alignements de séquences ainsi que l'analyse des alignements et les tests statistiques ont été réalisés de manière automatisée. Néanmoins, chaque transcrit a été traité de manière individuelle. La confirmation de l'hypothèse ainsi formulée à plus grande échelle

est indispensable pour valider le concept bioinformatique mais nécessite de profondes modifications dans l'approche même, les données utilisées ainsi que les scripts mis en œuvre. De plus, l'ouverture du concept d'infidélité de transcription aux évènements de délétions et d'insertions de bases sera illustrée dans la troisième partie de ce manuscrit.

3 Extension du procédé d'analyse bioinformatique au transcriptome entier

L'étude a été réalisée en 2007 et actualisée en 2008. Seuls les résultats les plus récents sont présentés.

3.1 Données et outils utilisés

3.1.1 Optimisation de la démarche bioinformatique

Dans l'étude préliminaire, l'ensemble des ESTs était aligné à chacun des transcrits successivement. Tout alignement présentant plus de 90% d'identité était retenu. L'inconvénient de cette méthode est qu'une EST peut s'aligner avec plusieurs transcrits (homologues, pseudogènes, …).

Pour pallier cette situation, nous avons inversé la démarche de l'alignement. Chaque EST est alignée à l'ensemble des séquences de référence ARN et seul le meilleur alignement est retenu. L'alignement retenu doit présenter les caractéristiques suivantes :

- Pourcentage d'identité minimum de 90%,
- E-value la plus petite de l'ensemble des alignements.

La E-value indique la probabilité qu'un alignement de même score soit trouvé « au hasard » dans la banque de données. Le score est calculé à partir du nombre d'identités, de substitutions et de délétions de l'alignement. Il est important de noter que nous avons choisi de ne pas pénaliser l'introduction de délétions dans l'alignement. Ce choix sera argumenté dans la section 3.4.4.

Toutes les séquences ESTs sont traitées successivement de cette manière.

<u>Différentes étapes de l'analyse :</u>

- ✓ Création de 2 banques de données, la première contenant l'ensemble des séquences de référence ARN et la seconde l'ensemble des ESTs humaines,
- ✓ Alignement MegaBLAST paramétré de la manière suivante :
    - « query » = banque d'ESTs,
    - « database » = banque de séquences de référence ARN,
    - Paramètres par défaut, excepté :

        -q = -2, pénalité d'un mismatch,

        -b = 1, nombre de sorties (seul le meilleur alignement est ainsi retenu),

-p = 90, pourcentage d'identité minimum entre la RefSeq et l'EST,

-W = 16, taille du mot d'encrage,

- ✓ Création de deux ensembles d'alignements en fonction de l'origine cancéreuse ou non du tissu à partir duquel l'EST a été extraite,
- ✓ Création de fichiers pour chaque transcrit représenté dans la banque ESTs. Pour chaque transcrit, création d'un premier fichier contenant les alignements entre le transcrit et les ESTs d'origine normale et d'un second fichier contenant les alignements entre le transcrit et les ESTs d'origine cancéreuse.
- ✓ Lecture des alignements pour chaque transcrit successivement et création de fichiers contenant la nature de la base portée par chaque EST pour l'ensemble des positions du transcrit.

Représentation graphique de la démarche bioinformatique mise en place pour analyser le génome entier :

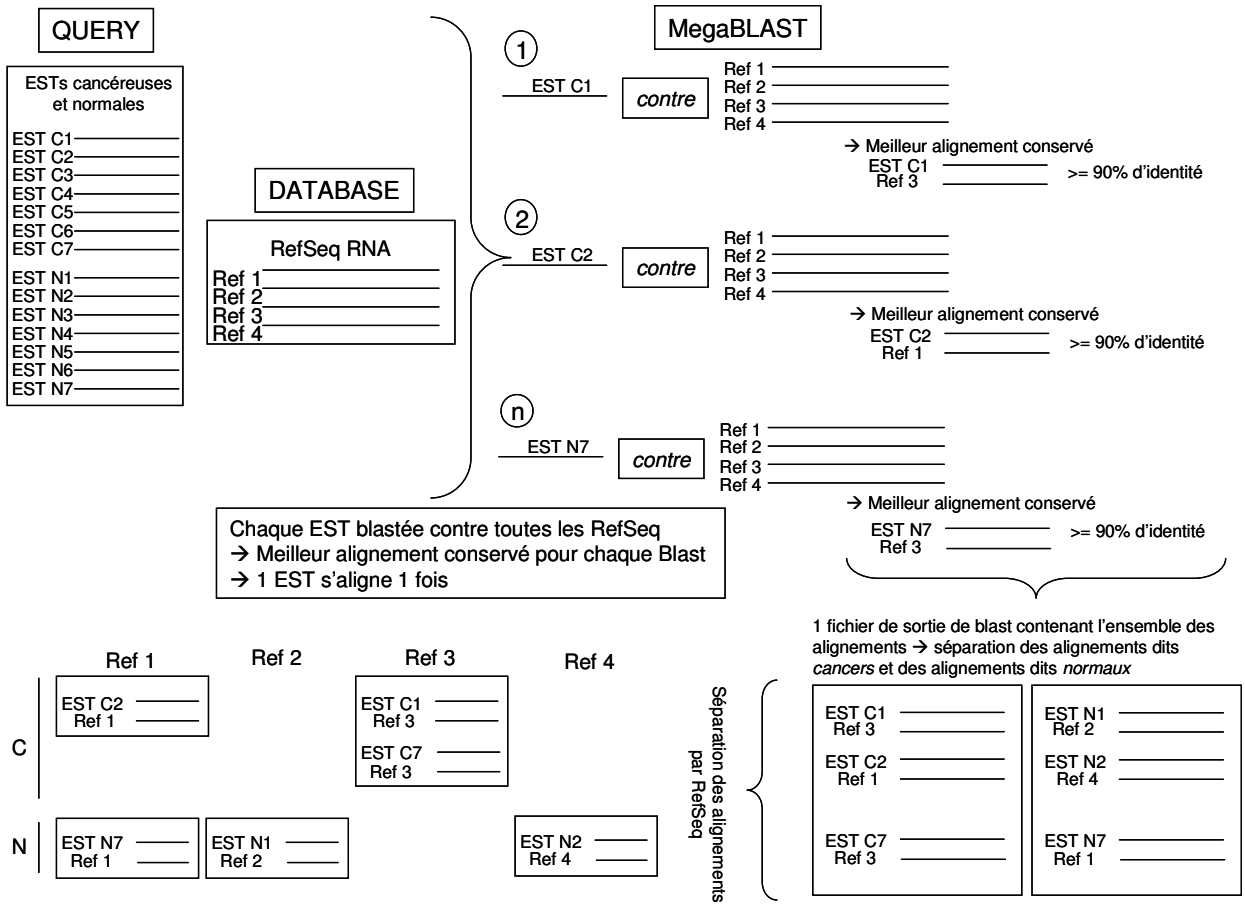


Figure 17 : Approche bioinformatique de l'analyse du génome entier.

3.1.2 Les séquences de référence ARN

La banque de séquences RefSeq du NCBI [134], sous-division de Genbank [128], a été choisie comme banque de référence car il existe un système de validation des séquences. RefSeq contient des séquences validées, annotées, non redondantes, et reflète la synthèse des connaissances actuelles pour chacune des entrées.

L'ensemble des séquences de référence ARN du NCBI a été téléchargé le 18 juillet 2008. Les RefSeq exploitées sont des séquences ARN de différentes natures :

- Séquences **NM** : ARNm

  25.733 séquences,

- Séquences **NR** : transcrits non codants (ARN structuraux, pseudogènes transcrits)

  1.123 séquences,

- Séquences **XM** : modèles d'ARNm issus de procédés automatisés d'annotation du génome

  12.401 séquences,

- Séquences **XR** : modèles de transcrits non codants issus également de procédés automatisés d'annotation du génome

  6.306 séquences.

L'ensemble de données initial contient 45.563 transcrits.

3.1.3 Mise à jour des ESTs utilisées

Les ESTs humaines utilisées ont été téléchargées depuis le site du NCBI le 3 juillet 2008.

Seules les ESTs déposées après le 1er janvier 2000 sont prises en compte.

L'ensemble de données initial contient 6.446.485 ESTs, dont 2.608.711 séquences identifiées comme étant issues de tissus cancéreux et 2.985.560 séquences issues de tissus non cancéreux.

3.1.4 Mise à jour du BLAST

La dernière mise à jour de l'outil MegaBLAST (version 2.2.18) a été utilisée.

Nous avons choisi, lors de la mise à jour 2008, de réaliser un BLAST unique et de trier par la suite les alignements en fonction de l'origine cancéreuse ou non des ESTs.

3.1.5 Différents évènements d'infidélité de transcription

3 types d'évènements sont définis :

- ✓ Substitution de base : l'EST porte une base de nature différente de celle définie par la séquence de référence,
- ✓ Délétion : la référence porte une base qui est absente de l'EST ; pour prolonger l'alignement, MegaBLAST crée alors un «gap» artificiel sur l'EST.
- ✓ Insertion : une base est insérée entre deux positions consécutives sur la séquence de référence ; pour prolonger l'alignement, MegaBLAST crée un «gap» dans la séquence de référence.

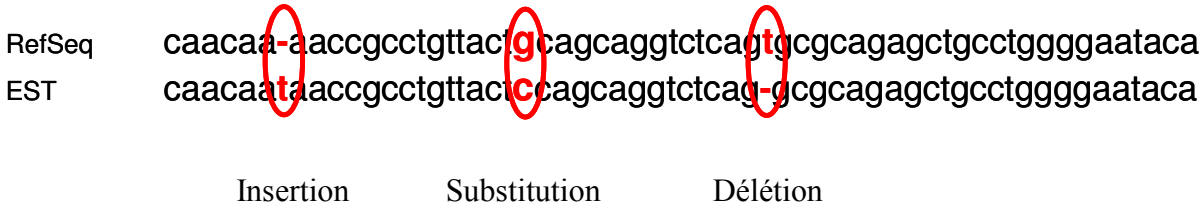


Figure 18 : Représentation des 3 types d'évènements étudiés.

3.1.6 Mise en place d'un second test statistique

Le test de proportions précédemment présenté est très conservateur et puissant mais ses contraintes sont drastiques (effectif total >70 et effectifs théoriques >5) et le nombre de positions analysables est donc restreint.

Il existe en particulier des positions pour lesquelles le test de proportions n'est pas applicable mais qui sont néanmoins des positions d'intérêt. Le tableau de contingence ci-dessous en est un exemple :

	Base de référence	Base infidèle
ESTs normales	**200**	**0**
ESTs cancéreuses	**200**	**10**

Table 9 : Exemple de tableau de contingence ne respectant pas la contrainte du test de proportions concernant les effectifs théoriques.

Les effectifs théoriques sont les suivants :

	Base de référence	Base infidèle
ESTs normales	**195.12**	**4.88**
ESTs cancéreuses	**204.88**	**5.12**

Table 10 : Effectifs théoriques correspondant à la table 9.

Un effectif théorique étant inférieur à 5, les contraintes du test de proportions ne sont pas respectées.

Pour remédier à cette situation, nous avons mis en place, pour chaque position, deux statistiques successives :

- Si les contraintes du test de proportions sont respectées, alors application du test de proportions,
- Sinon, alors application du test exact de Fisher. Les conditions d'application de ce test sont définies arbitrairement de la manière suivante :

  Effectif total > 70

  Effectifs théoriques ≥ 2.

  Le test de Fisher n'impose pas de contraintes. Travaillant sur des données contenant, par définition, des erreurs de séquençage, nous avons néanmoins décidé d'appliquer des contraintes d'effectifs.

Les différentes présentations du travail réalisées au cours de l'étude préliminaire nous ont appris que la critique récurrente des biologistes et bioinformaticiens portait sur le manque de fiabilité des séquences ESTs. Le principe du test statistique repose sur une comparaison de deux ensembles, c'est pourquoi les erreurs de séquençage ne peuvent pas influencer les résultats de l'analyse statistique. Nous savons par ailleurs que l'étude préliminaire résiste à l'application de filtres (trop) drastiques. Il est néanmoins nécessaire de prendre ces réserves en considération, d'où la mise en place de différents filtres dans cette étude à grande échelle.

3.2 Filtres appliqués

3.2.1 Filtre d'alignements

Forts des connaissances acquises lors de l'étude préliminaire sur 17 transcrits, les 50 premiers et derniers éléments de chaque alignement sont éliminés de l'analyse. Par conséquent, seuls les alignements de plus de 100 éléments sont conservés. Par ailleurs, comme précédemment, les alignements couvrant moins de 70% de la longueur totale de l'EST sont éliminés.

3.2.2 Filtre de positions

Les alignements sont lus de manière brute ou avec un filtre défini de la manière suivante :

Une base n'est lue sur l'EST et comptabilisée qu'à la condition que les 10 bases précédant et les 10 bases suivant l'événement soient parfaitement alignées avec la séquence de référence.

Ce filtre sera nommé filtre -10/+10.

Des recherches automatiques de SNPs réalisées à partir des séquences ESTs sont pratiquées sur le même principe, avec des filtres allant de 5 à 20 bases parfaitement alignées de part et d'autre de chaque événement [112,115]. Ce filtre est justifié par le fait que les regroupements de mésappariements se retrouvent souvent au niveau de régions où le séquençage est de faible qualité [112].

3.2.2.1 Substitutions

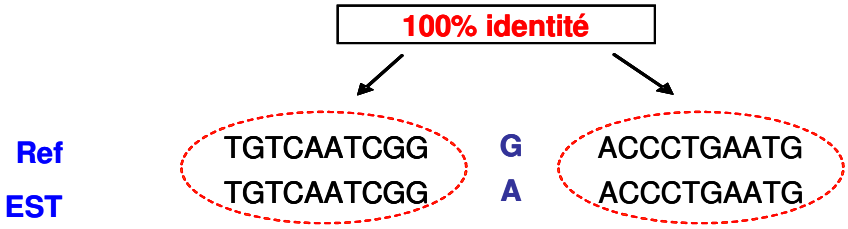


Figure 19 : Filtre appliqué aux évènements de substitutions.

Le cas des substitutions est le cas le plus simple ; il ne nécessite pas de traitement supplémentaire.

3.2.2.2 Délétions

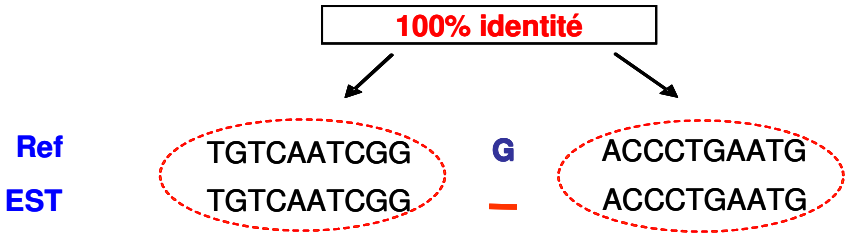


Figure 20 : Filtre appliqué aux évènements de délétions.

Au sein d'une répétition de bases (ou n-uplet) se pose le problème de la localisation de la délétion. Dans l'exemple ci-dessous, la première délétion peut survenir au niveau de la 1ère, 2ème ou 3ème thymine du triplet TTT.

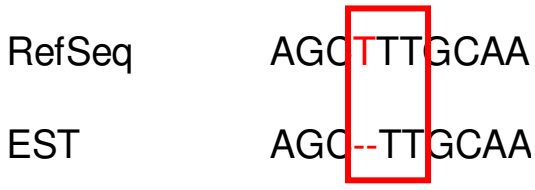


Figure 21 : Problème d'une délétion survenant au sein d'une répétition de bases.

Une difficulté supplémentaire imputable au MegaBLAST réside dans le fait que l'outil positionne la délétion en fonction du point d'encrage de l'alignement. Deux situations peuvent ainsi être rencontrées au sein du même alignement :

```
Query  1   CATGCATGCAAAAATGCATGCATGCATGCACGTACGTACACCACATGACGTTGCAGTCAG  60
           ||||||||||||||  ||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1   CATGCATGCAAAA-TGCATGCATGCATGCACGTACGTACACCACATGACGTTGCAGTCAG  59

Query  61  TGCACACGTACGCATGCAAAACGATGCAGT  90
           |||||||||||||||| ||||||||||||
Sbjct  60  TGCACACGTACGCATGC-AAACGATGCAGT  88
```

Figure 22 : La délétion peut être localisée en début ou en fin de la répétition de bases.

Ne pouvant déterminer la localisation réelle des délétions, chaque délétion est positionnée arbitrairement sur le dernier nucléotide du n-uplet lors de la lecture de l'alignement. Cette décision arbitraire est sans impact biologique. En effet, quelle que soit la position réelle de la délétion, son impact codant, *i.e.* le décalage du cadre de lecture, sera le même.

3.2.2.3 Insertions

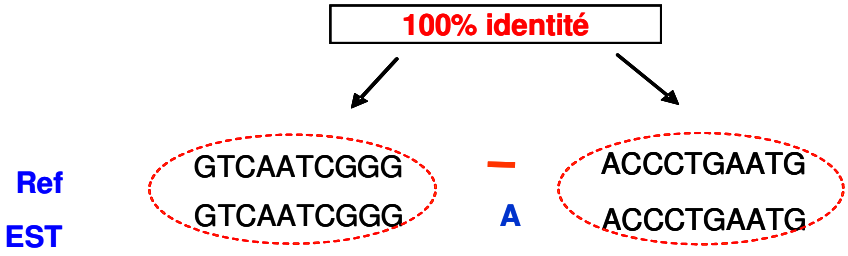


Figure 23 : Filtre appliqué aux évènements d'insertions.

Pour les mêmes raisons que celles évoquées dans le cas des délétions, la localisation vraie des insertions pose problème.

Deux cas sont distingués :

- ✓ Les insertions isolées : la base insérée est de nature différente de celle des bases flanquantes.

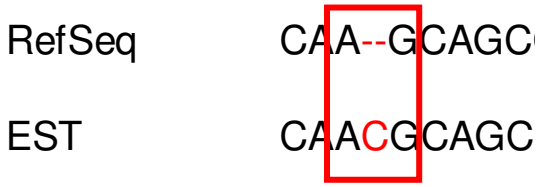


Figure 24 : Exemple d'insertion isolée.

Une position est définie par défaut par rapport à la position de la base sur l'ARN de la séquence de référence correspondante. Dans le cas d'une insertion, il n'existe pas de base correspondante sur l'ARN de référence. Une insertion isolée sera arbitrairement localisée sur la position de l'ARN de référence précédant cette insertion.

- ✓ Les insertions générant un n-uplet : la base insérée est de même nature que la base précédant ou suivant cette insertion.

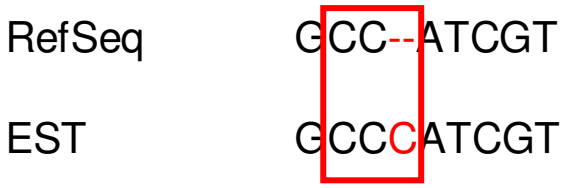


Figure 25 : Exemple d'insertion liée à une répétition de bases.

Une insertion générant un n-uplet sera référencée au niveau de la dernière base du n-uplet sur l'ARN de référence.

Les filtres appliqués sont donc différents de ceux de l'étude préliminaire. La réalisation des tests statistiques a donc été adaptée.

3.3 Application des tests statistiques

La lecture des alignements est plus complexe que dans le cas de l'étude préliminaire. Les tableaux de données pour chaque transcrit sont complétés de la manière suivante :

ESTs d'origine cancéreuse										ESTs d'origine non cancéreuse								
A	T	C	G	del	insA	insT	insC	insG	RefSeq	A	T	C	G	del	insA	insT	insC	insG
1	0	992	19	0	0	0	0	0	C	0	0	709	1	0	0	0	0	0
0	0	1013	0	0	0	0	10	0	C	0	0	709	0	0	0	0	0	0
0	985	27	0	0	0	0	0	0	T	0	692	18	0	0	0	0	0	0
0	0	1013	0	0	0	0	0	0	C	0	0	707	0	0	0	0	0	0
0	1017	0	0	0	0	0	0	0	T	0	707	0	0	0	0	0	0	0
1	1016	1	0	38	0	0	0	0	T	0	705	0	0	2	0	1	0	0
1	0	1014	0	0	0	0	0	0	C	0	0	706	0	0	0	0	0	0
0	0	1016	0	0	0	0	0	0	C	0	0	704	0	0	0	0	0	0
988	2	25	1	0	2	0	0	0	A	701	0	3	0	0	0	0	0	0

Table 11 :Résultat de l'exploitation de la sortie de Blast.

Les tableaux de contingence sont définis de la manière suivante :

- ✓ Substitutions

ESTs normales identiques à RefSeq	ESTs normales substituées
ESTs cancéreuses identiques à RefSeq	ESTs cancéreuses substituées

Table 12 :Tableau de contingence pour les substitutions.

- ✓ Délétions

ESTs normales identiques à RefSeq	ESTs normales portant une délétion
ESTs cancéreuses identiques à RefSeq	ESTs cancéreuses portant une délétion

Table 13 :Tableau de contingence pour les délétions.

- ✓ Insertions

Pour chaque position, les tests statistiques sont appliqués aux insertions dites « isolées », aux insertions dites « n-uplet » et enfin à la somme des insertions.

ESTs normales identiques à RefSeq	ESTs normales portant une insertion
ESTs cancéreuses identiques à RefSeq	ESTs cancéreuses portant une insertion

Table 14 :Tableau de contingence pour les insertions.

L'étude des ESTs représentatives de l'ensemble du transcriptome humain peut être représentée de la manière suivante :

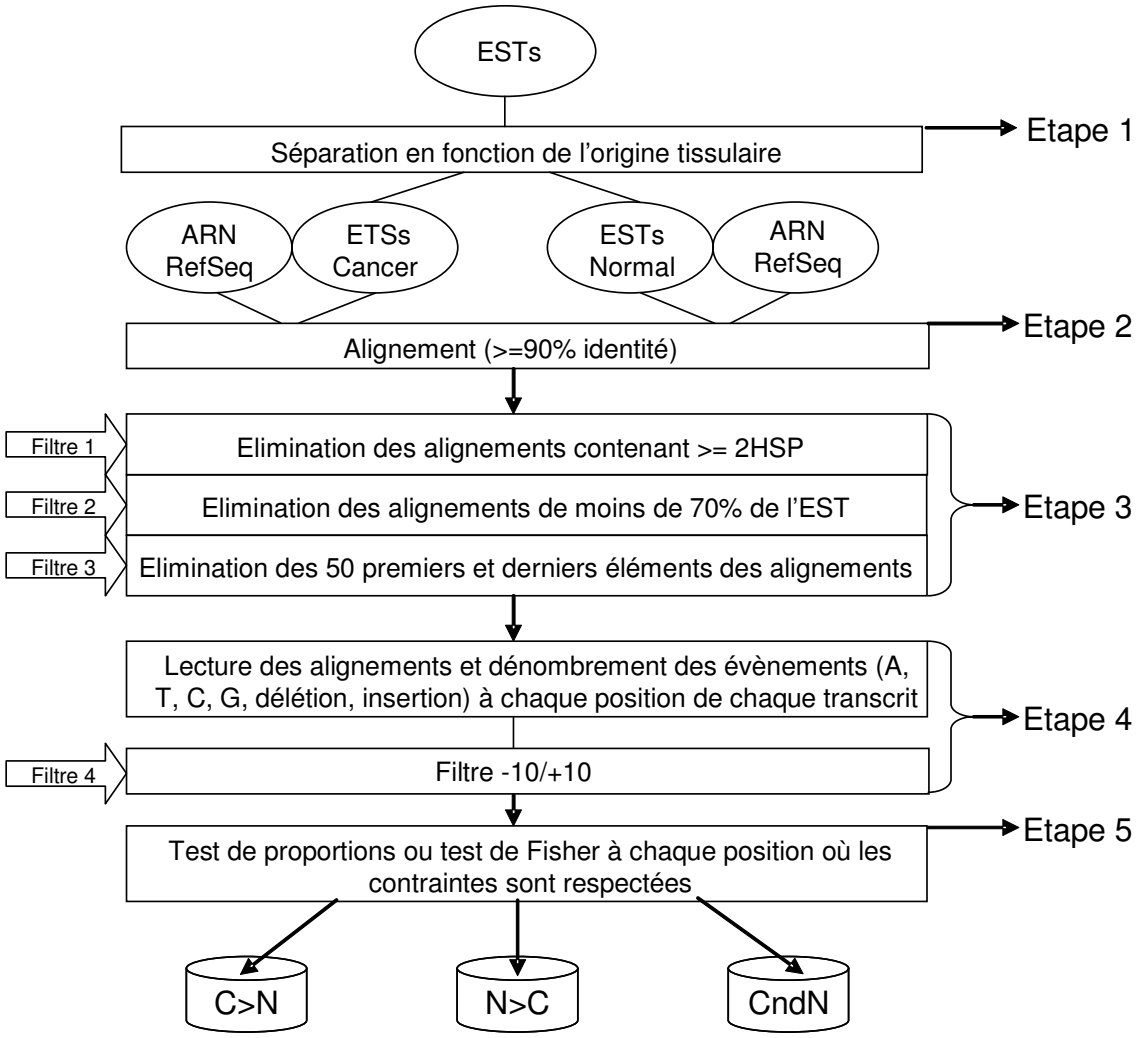


Figure 26 : Différentes étapes de l'étude du transcriptome.

3.4 Résultats

Une distinction est faite entre les évènements dits simples, *i.e.* impliquant une seule base, et les évènements dits multiples impliquant plusieurs bases.

3.4.1 Évènements simples

3.4.1.1 Effectifs

L'ensemble des transcrits étudiés permet l'analyse de près de 90 millions de positions d'ARN. Les effectifs des positions respectant les contraintes des tests statistiques (Chi² et Fisher exact) sont présentés dans le tableau ci-dessous. Les positions pour lesquelles un test statistique est réalisable représentent seulement un mince sous-ensemble du nombre total de positions observables. En effet, parmi l'ensemble des transcrits analysés, seules 4% des positions

satisfont la première contrainte d'effectif ; en d'autres termes, les positions contre lesquelles plus de 70 ESTs sont alignées ne représentent que 4% de l'ensemble des positions des ARN. Cette analyse est donc limitée à l'observation d'un sous-ensemble de transcrits contre lesquels un nombre suffisant d'ESTs s'aligne. En règle générale, il est admis que la proportion d'un clone ADNc dans une librairie est équivalente à la proportion de l'ARN correspondant dans la cellule. Par conséquent, la présente analyse ne permet de rendre compte que des évènements affectant les transcrits moyennement à abondamment exprimés, mais ne permet pas de traiter les transcrits rares.

Nombre total de positions analysées	87 012 313

Substitutions		
Effectifs > 70 ESTs	3 593 436	
	brut	filtre
Test du Chi² : effectifs théoriques >5	24709	6262
Test de Fisher : effectifs théoriques >1	66704	15336

Délétions (Gap)		
Effectifs > 70 ESTs	3 583 371	
	brut	filtre
Test du Chi² : effectifs théoriques >5	17068	4108
Test de Fisher : effectifs théoriques >1	38948	9065

Insertions isolées (InsI)		
Effectifs > 70 ESTs	3 577 909	
	brut	filtre
Test du Chi² : effectifs théoriques >5	5078	776
Test de Fisher : effectifs théoriques >1	15668	2992

Insertions n-uplet (InsN)		
Effectifs > 70 ESTs	3 583 008	
	brut	filtre
Test du Chi² : effectifs théoriques >5	12358	1787
Test de Fisher : effectifs théoriques >1	34375	5495

Insertions totales (InsT)		
Effectifs > 70 ESTs	3 589 041	
	brut	filtre
Test du Chi² : effectifs théoriques >5	20121	3003
Test de Fisher : effectifs théoriques >1	57161	9624

Table 15 : Nombre de positions analysables pour les différents types d'évènements, avec ou sans application du filtre.

Remarque : le test de Fisher n'est réalisé qu'à la condition que les contraintes du test du Chi² ne soient pas respectées. Les positions dites Fisher sont par conséquent différentes des positions dites Chi².

Seules 0,2% de l'ensemble des positions ont réuni les conditions d'application d'un test statistique.

3.4.1.2 Résultats des tests statistiques

- Données brutes

Le nombre de positions pour lesquelles la proportion d'ESTs d'origine cancéreuse est supérieure, inférieure ou non différente de la proportion d'ESTs issues de tissus non tumoraux est présenté ci-dessous pour les différents types d'évènements d'infidélité de transcription précédemment présentés :

Substitutions		C>N	LBE	N>C	LBE	CndN		(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	8618	824	2632	1647	13459		3,3	7,9
	Fisher	7340	3113	2425	4732	56939		3,0	ND
	somme	15958	3937	5057	6379	70398		3,2	ND

Délétions		C>N	LBE	N>C	LBE	CndN		(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	8627	431	1710	1276	6731		5,0	18,9
	Fisher	7294	1544	1769	2937	29885		4,1	ND
	somme	15921	1975	3479	4213	36616		4,6	ND

Insertions I		C>N	LBE	N>C	LBE	CndN		(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	2414	130	357	378	2307		6,8	ND
	Fisher	2783	565	438	1246	12447		6,4	ND
	somme	5197	695	795	1624	14754		6,5	ND

Insertions N		C>N	LBE	N>C	LBE	CndN		(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	4330	369	832	867	7196		5,2	ND
	Fisher	3623	1529	905	2525	29847		4,0	ND
	somme	7953	1898	1737	3392	37043		4,6	ND

Insertions T		C>N	LBE	N>C	LBE	CndN		(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	7867	552	1253	1460	11001		6,3	ND
	Fisher	7108	2393	1337	4319	48716		5,3	ND
	somme	14975	2945	2590	5779	59717		5,8	ND

Table 16 : Exploitation des données non filtrées. Nombre de positions présentant un test significatif ou non pour chaque type d'événement.

ND signifie non déterminé (dénominateur nul).

La première conclusion établie sur ces données est que l'on confirme, sur l'ensemble du transcriptome, les données obtenues sur les 17 transcrits de l'étude préliminaire. En effet, les substitutions C>N sont détectées par le test du Chi² trois fois plus fréquemment que les substitutions N>C.

L'hétérogénéité des ARN issus de tissus cancéreux et décrite sur 17 gènes est donc confirmée à grande échelle.

La seconde conclusion est que les évènements de délétions et d'insertions sont détectés avec la même intensité que les substitutions.

Le test de Fisher permet d'accroître fortement le nombre de positions testées.

Il est intéressant de noter que, quel que soit le type d'événement, le nombre de tests de Fisher N>C est inférieur à l'estimation du nombre de faux positifs et le nombre de tests du Chi² N>C est du même ordre de grandeur que le LBE. Ainsi, des positions N>C sont détectées, mais elles ne sont pas supérieures au bruit statistique.

- Données filtrées

Les données filtrées sont obtenues après application du filtre -10/+10 décrit dans le paragraphe 3.2.2.

Substitutions		C>N	LBE	N>C	LBE	CndN		(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	2013	243	944	383	3305		2,1	3,2
	Fisher	1725	769	869	1027	12742		2,0	ND
	somme	3738	1012	1813	1410	16047		2,1	6,8

Délétions		C>N	LBE	N>C	LBE	CndN		(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	2912	56	212	355	984		13,7	ND
	Fisher	2733	260	272	759	6060		10,0	ND
	somme	5645	316	484	1114	7044		11,7	ND

Insertions I		C>N	LBE	N>C	LBE	CndN		(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	381	24	74	54	321		5,1	17,9
	Fisher	584	113	71	232	2337		8,2	ND
	somme	965	137	145	286	2658		6,7	ND

Insertions N		C>N	LBE	N>C	LBE	CndN		(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	676	53	162	126	949		4,2	17,3
	Fisher	593	247	178	398	4724		3,3	ND
	somme	1269	300	340	524	5673		3,7	ND

Insertions T		C>N	LBE	N>C	LBE	CndN		(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	1250	85	241	215	1512		5,2	44,8
	Fisher	1316	402	260	722	8048		5,1	ND
	somme	2566	487	501	937	9560		5,1	ND

Table 17 : Exploitation des données filtrées. Nombre de positions présentant un test significatif ou non pour chaque type d'événement.

L'impact du filtre est notable. Le tableau ci-dessous montre l'impact du filtre -10/+10 :

		C>N	N>C
Substitutions	Chi²	4,28	2,79
	Fisher	4,26	2,79
Délétions	Chi²	2,96	8,07
	Fisher	2,67	6,50
Insertions T	Chi²	6,29	5,20
	Fisher	5,40	5,14

Table 18 : rapport entre les positions issues de l'analyse brute et les positions issues de l'analyse après filtre.

L'application du filtre -10/+10 ne modifie pas les conclusions : l'hétérogénéité est plus importante au sein des ESTs cancéreuses que des ESTs non cancéreuses.

L'impact du filtre est important sur les insertions, pour lesquelles le nombre de positions statistiquement significatives chute d'un facteur six.

L'impact du filtre est particulièrement intéressant sur les délétions puisque le filtre provoque une baisse des positions C>N d'un facteur trois parallèlement à une baisse des positions N>C d'un facteur sept. Le différentiel entre le nombre de positions C>N et N>C est donc amplifié par l'application du filtre -10/+10. Ainsi, les délétions sont les évènements pour lesquels l'hétérogénéité accrue dans les séquences d'origine cancéreuse est la plus grande.

La figure 27 donne une représentation des résultats après application du filtre -10/+10 :

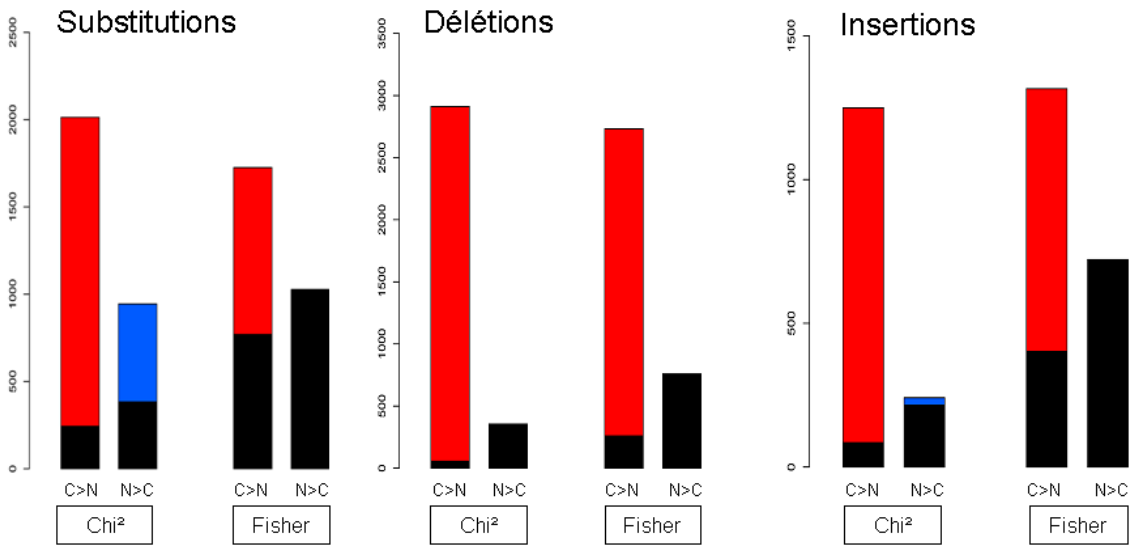


Figure 27 : Représentation du nombre de positions C>N et N>C pour les 3 types d'évènements et les deux tests utilisés.

La couleur rouge représente les positions C>N, la bleue les N>C. Les parties noires représentent le LBE. Le même code couleur sera appliqué aux figures 28 à 30.

3.4.2 Évènements multiples

Un événement multiple est défini comme une délétion ou une insertion de plusieurs bases consécutives.

3.4.2.1 Modification des paramètres du MegaBLAST

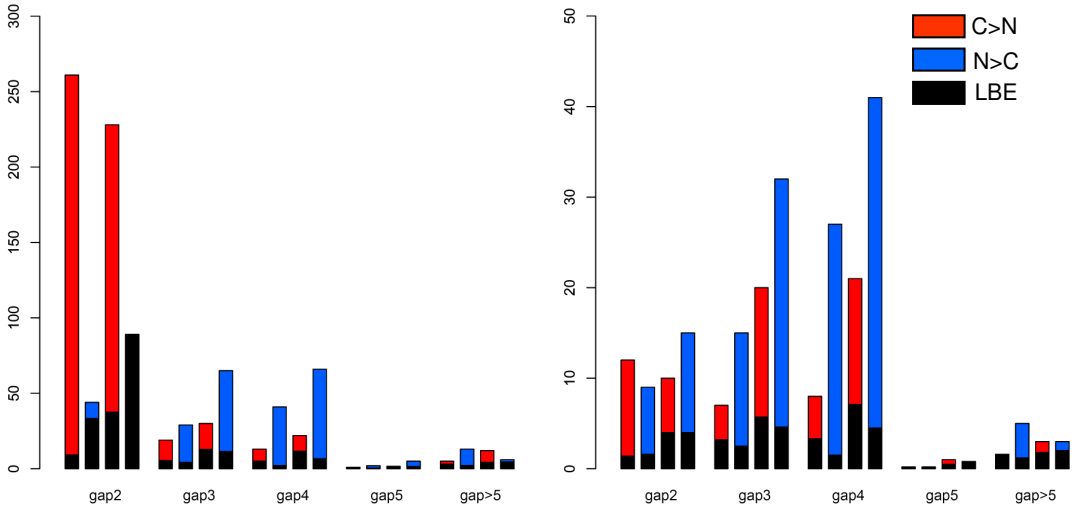
Les paramètres d'alignement appliqués précédemment ne sont pas adaptés à l'étude des délétions et insertions multiples, dans la mesure où la création et l'extension des gaps n'était pas pénalisée. La création et l'extension d'un gap sont ici pénalisées (-G 1, -E 1).

3.4.2.2 Résultats

Les délétions et insertions multiples dont la proportion est augmentée dans les tissus cancéreux par rapport aux tissus normaux (ou inversement) existent.

Le filtre -10/+10 utilisé dans cette partie a dû être adapté et correspond à un alignement parfait des 10 bases précédant et des 10 bases suivant la délétion ou l'insertion multiple.

- ✓ Délétions multiples :



Données brutes Données filtrées

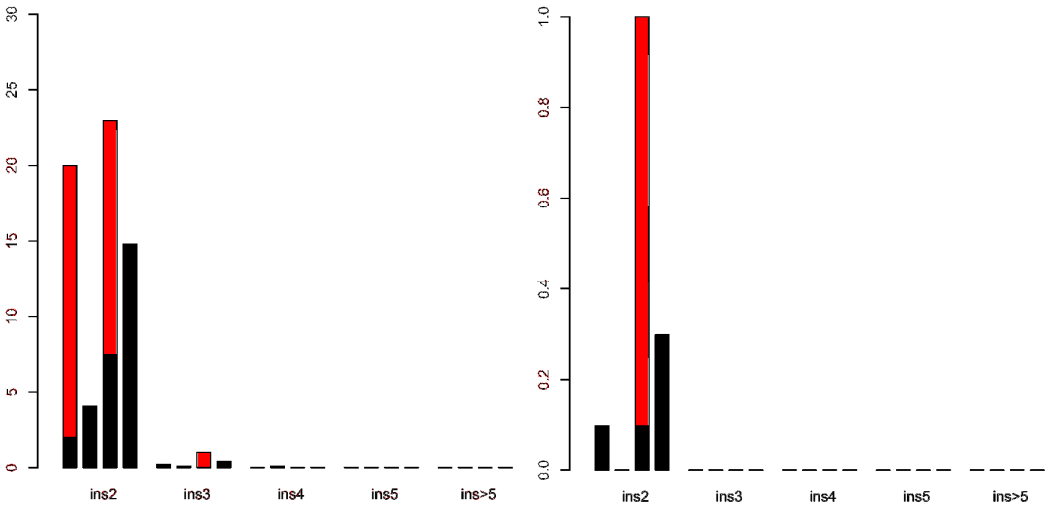
Figure 28 : Délétions. Représentation du nombre de positions C>N et N>C en fonction du nombre de bases délétées.

Les résultats de l'étude des délétions multiples dépendent du filtre appliqué.

Avant application du filtre, les délétions C>N de deux bases issues des tests de proportions et de Fisher sont nombreuses (500) et largement supérieures au LBE ; les délétions N>C sont inférieures au bruit statistique. Les délétions de trois et quatre bases sont au contraire des délétions majoritairement N>C. Les délétions de cinq bases et plus sont anecdotiques.

Après application du filtre, le nombre de délétions C>N de deux bases chute. Les délétions N>C de trois et quatre bases sont nettement moins sensibles à l'application du filtre et restent à la fois supérieures au LBE et supérieures au nombre de délétions C>N.

✓ Insertions N

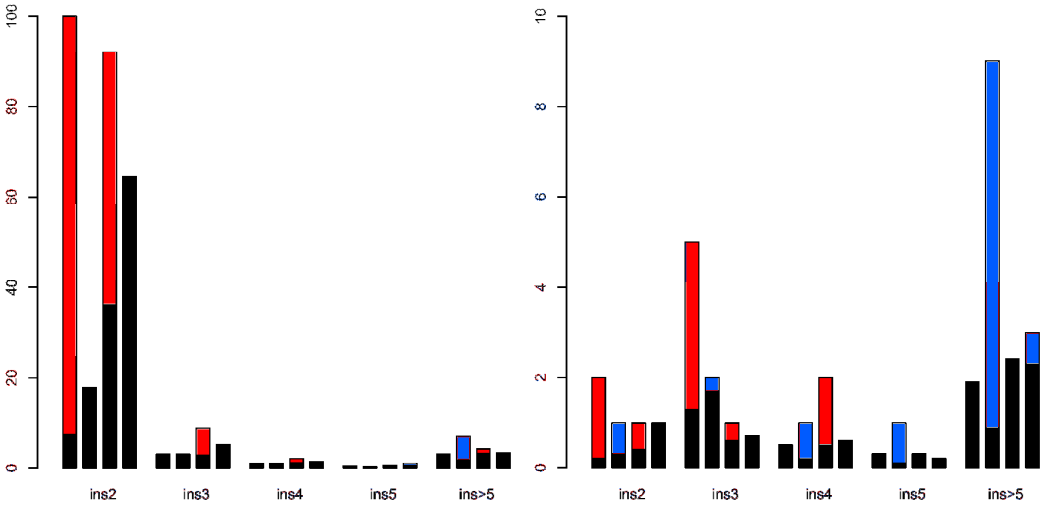


Données brutes Données filtrées

Figure 29 : Insertions au sein d'une répétition de bases. Représentation du nombre de positions C>N et N>C en fonction du nombre de bases insérées.

Les insertions de deux bases de même nature existent et sont exclusivement des évènements C>N. Ces insertions ne résistent pas à l'application du filtre.

✓ Insertions I



Données brutes Données filtrées

Figure 30 : Insertions isolées. Représentation du nombre de positions C>N et N>C en fonction du nombre de bases insérées.

Les résultats de l'étude des insertions multiples de nature hétérogène suivent le même profil que celui des délétions multiples. Ainsi, l'analyse réalisée avant application du filtre révèle un nombre important d'insertions C>N de deux bases de nature différente ; ces insertions ne résistent pas à l'application du filtre -10/+10.
Les délétions et insertions de plus de 20 bases ne sont pas traitées dans la mesure où de tels évènements sont anecdotiques et peuvent vraisemblablement être attribués à des épissages alternatifs non encore référencés.

Les résultats de l'étude préliminaire sont confirmés à grande échelle : il existe une hétérogénéité accrue portée par les ARN issus de tissus cancéreux par rapport aux ARN issus de tissus sains. Il s'est avéré nécessaire alors de vérifier l'importance du contexte local d'ADN et son adéquation avec la notion de bulle de transcription sur les données obtenues à l'échelle du transcriptome.

3.5 Étude du contexte d'ADN

3.5.1 Évènements d'infidélité de transcription

3.5.1.1 Règles de substitutions

L'étude préliminaire de 17 transcrits abondamment exprimés présentée en partie 2 a montré l'influence du contexte d'ADN sur la survenue de substitutions C>N, ainsi que l'existence de règles de remplacement de la base substituée par la base précédant ou suivant l'événement. L'objectif de cette partie est de vérifier ces résultats à plus grande échelle.
Pour cela, les substitutions C>N obtenues par application du test de proportions après application du filtre -10/+10 et situées dans la partie codante du transcrit sont analysées (n=2.013).
Ces substitutions peuvent être représentées de la manière suivante :

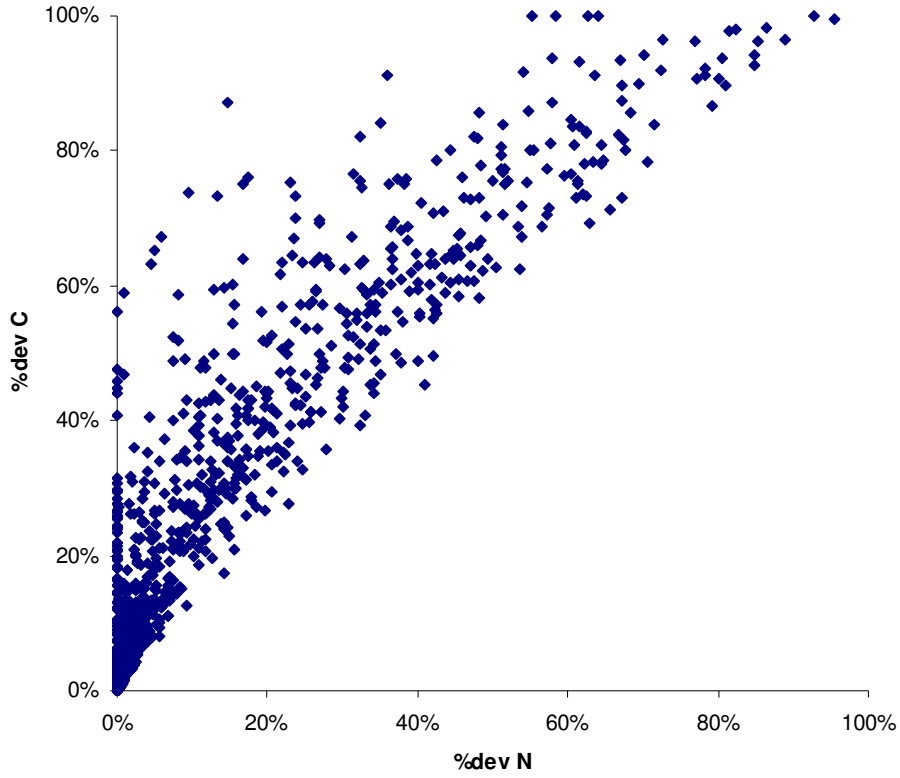


Figure 31 : Ensemble des substitutions C>N représentées par le pourcentage de déviation observé dans les ESTs cancéreuses en fonction du pourcentage de déviation observé dans les ESTs normales.

Les substitutions excédant 50% de déviation sont par définition douteuses, d'où la mise en place d'un filtre permettant d'éliminer les positions litigieuses. Une position litigieuse est définie comme étant une position pour laquelle la nature de la base RefSeq et la nature de la base majoritairement observée au sein des ESTs sont différentes (n=182). Ces positions sont éliminées de l'analyse.

Les positions de substitutions correspondant à des positions de SNP décrites dans dbSNP ont également été éliminées de l'analyse (n=281).

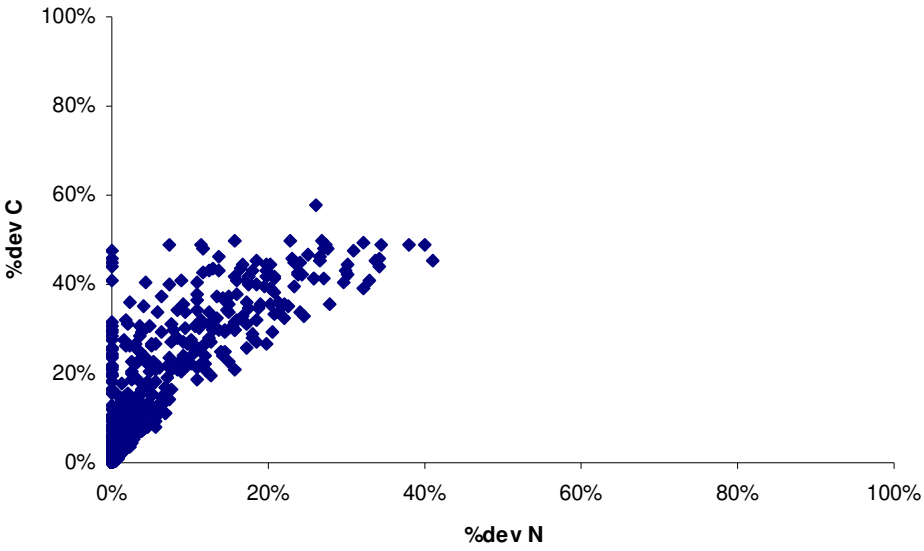


Figure 32 : Ensemble des substitutions C>N sélectionnées.

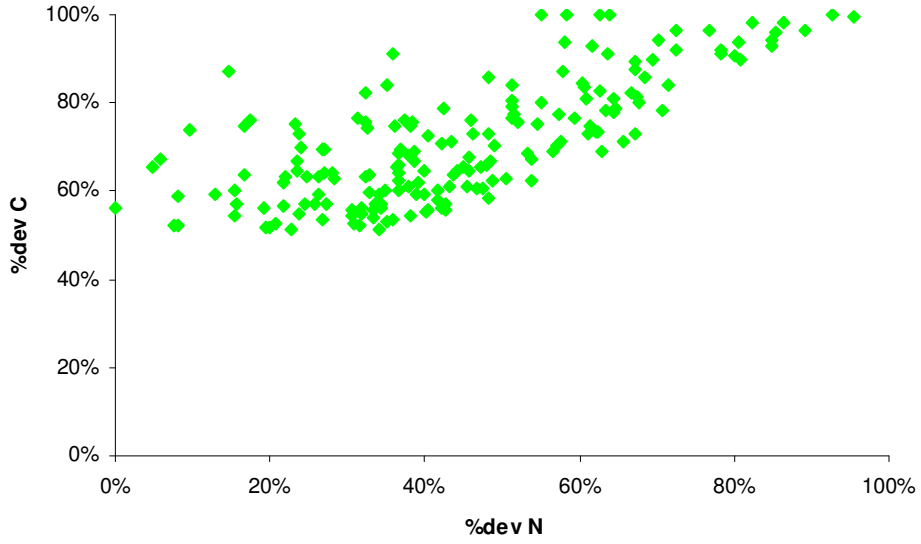


Figure 33 : Ensemble des substitutions C>N dites litigieuses.

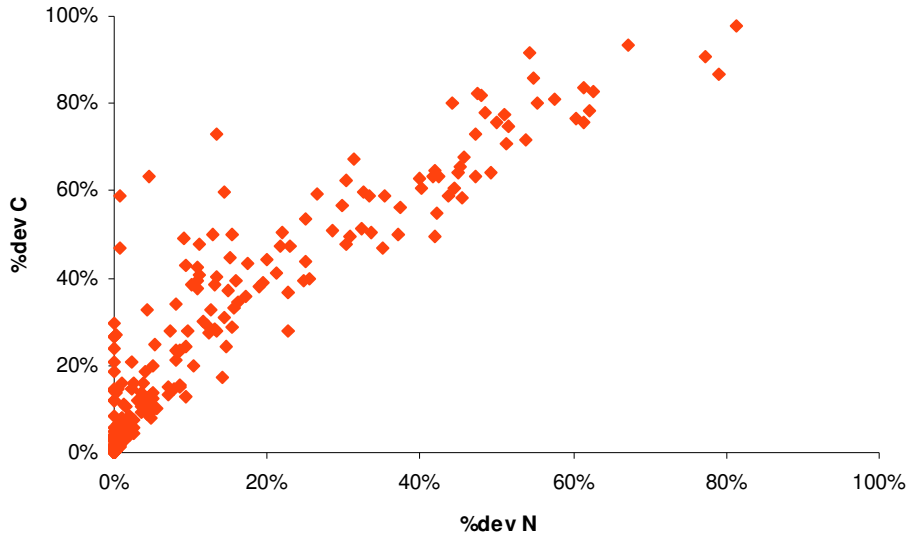


Figure 34 : Ensemble des substitutions C>N correspondant à un SNP.

Notons tout de même que parmi les 281 SNP décrits dans dbSNP et correspondant à une position de substitution C>N, seuls 124 SNPs ont été validés biologiquement. Les SNPs qui n'ont pas été validés peuvent être issus d'analyses à grande échelle d'ESTs. La variation peut dans ce cas être portée par l'ADN ou par l'ARN ; on ne peut donc pas conclure quant à la nature (SNP ou infidélité de transcription) de ces 124 positions. L'ensemble des 281 positions potentiellement portées par l'ADN a été écarté de l'étude.

La composition des bases affectées par les substitutions est la suivante :

A	**308**
T	**547**
C	**399**
G	**296**

Parmi les positions analysées, 432 sont en phase 1, 480 en phase 2 et 638 en phase 3. Il existe donc un déséquilibre important des substitutions C>N en faveur de la phase 3 (test d'ajustement du Chi² à la loi uniforme discrète, $p=2*10^{-10}$).
Or, les compositions en base des 3 phases sur les ARN étudiés présentant au moins une position d'intérêt, soit environ 10.000 RefSeq, sont différentes :

	phase1	phase2	phase3
A	26,3%	31,8%	18,3%
T	16,2%	26,5%	22,1%
C	24,3%	22,9%	29,8%
G	33,2%	18,8%	29,8%

Table 19 : Composition théorique en bases des trois phases.

Le calcul des effectifs théoriques doit donc s'effectuer en tenant compte de la phase. Par exemple, le nombre de A affectés attendu aléatoirement est égal à (0,263*432 + 0,318*480 + 0,183*638) soit 338.

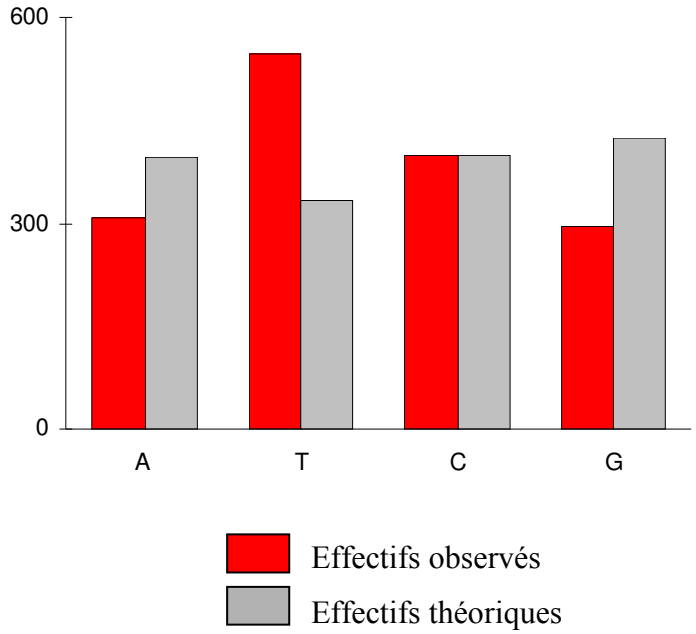


Figure 35 : Composition des bases affectées par les substitutions.

La base substituée de façon majoritaire est la base T ; ce résultat est cohérent avec l'étude préliminaire de 17 transcrits qui avait révélé que A et T étaient substitués majoritairement. La base substituée de façon minoritaire est la base G, tout comme dans l'étude préliminaire.
La composition des bases affectées est donc différente de la composition attendue par hasard (p-valeur du test du $Chi^2 = 5*10^{-38}$).

Pour chacune des bases affectées, la base de remplacement a été étudiée. N'ont été prises en compte que les substitutions pour lesquelles une base de remplacement majoritaire (>75%) peut être définie (n=820).

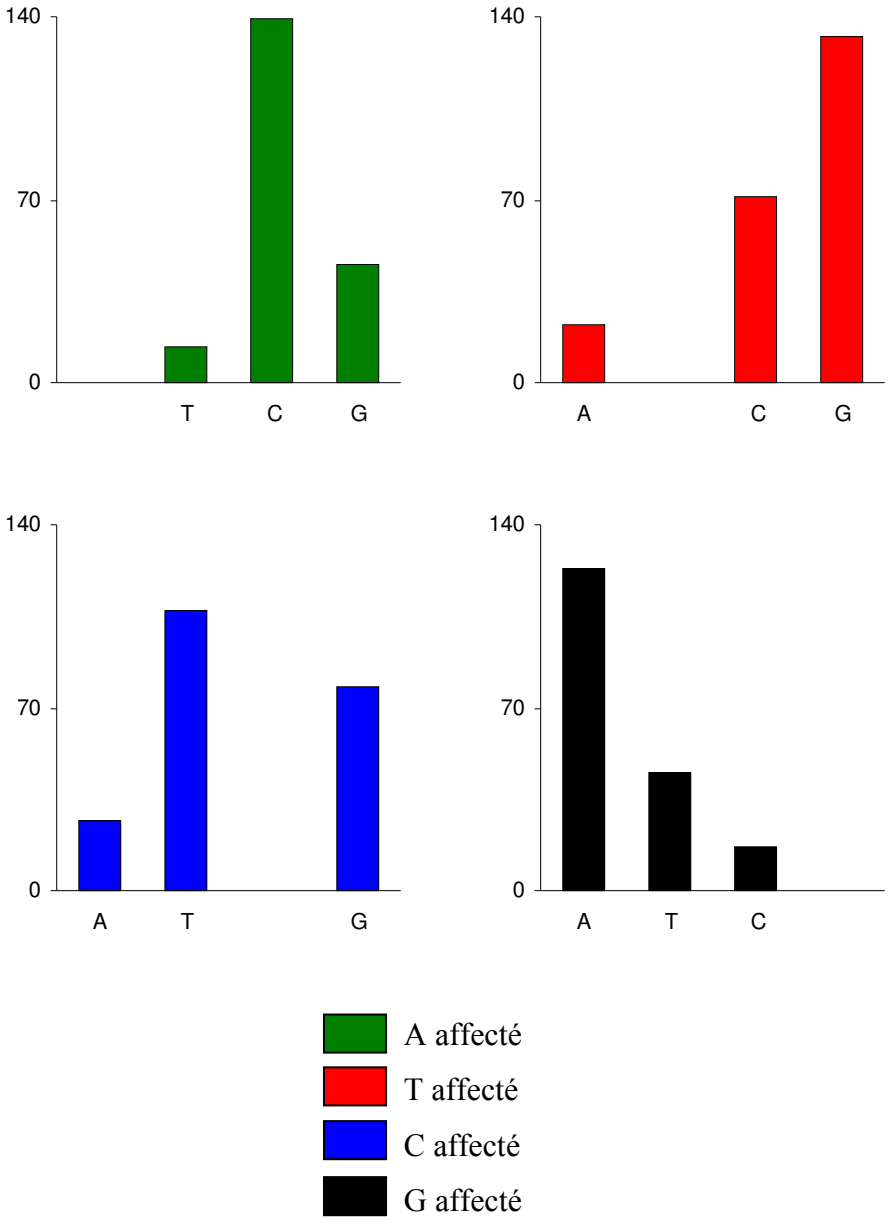


Figure 36 : Composition des bases de remplacement en fonction de la nature de la base affectée.

Les résultats établis sur 17 gènes sont donc retrouvés à grande échelle, *i.e.* A est préférentiellement remplacé par C (p-valeur du test d'ajustement à la loi uniforme = $1*10^{-28}$), T par G (p-valeur = $3*10^{-18}$), C par T et G (p-valeur = $8*10^{-11}$) et enfin G par A (p-valeur = $6*10^{-22}$).

Par ailleurs, l'étude à grande échelle confirme la première règle de remplacement établie lors de l'étude préliminaire et qui consiste à remplacer la base substituée par une base identique à celle qui vient d'être incorporée (b-1).

Le remplacement par la base suivant l'événement d'infidélité de transcription n'est pas confirmé.

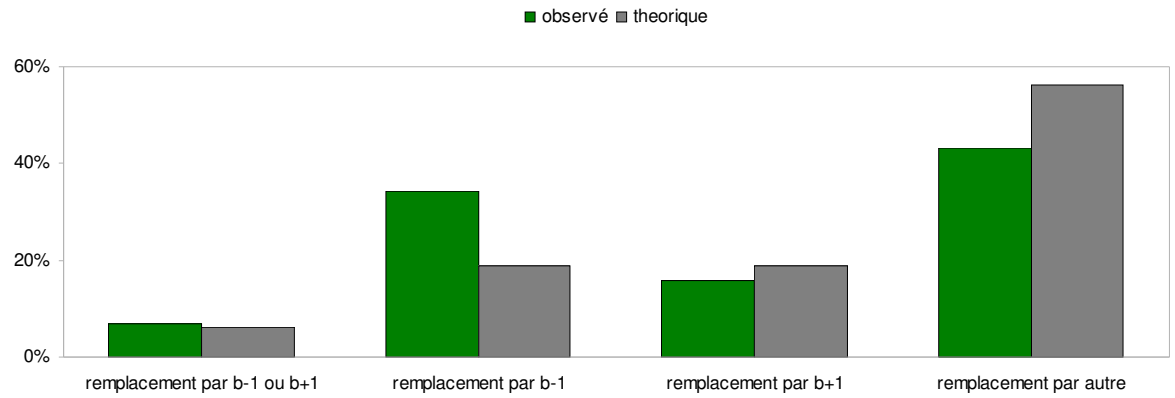


Figure 37 : Remplacement de la base affectée par la base précédant (b-1) ou suivant (b+1) la substitution C>N.

Pour finir, l'étude du contexte d'ADN flanquant les substitutions a été étudié de la même manière que dans le paragraphe 2.4.

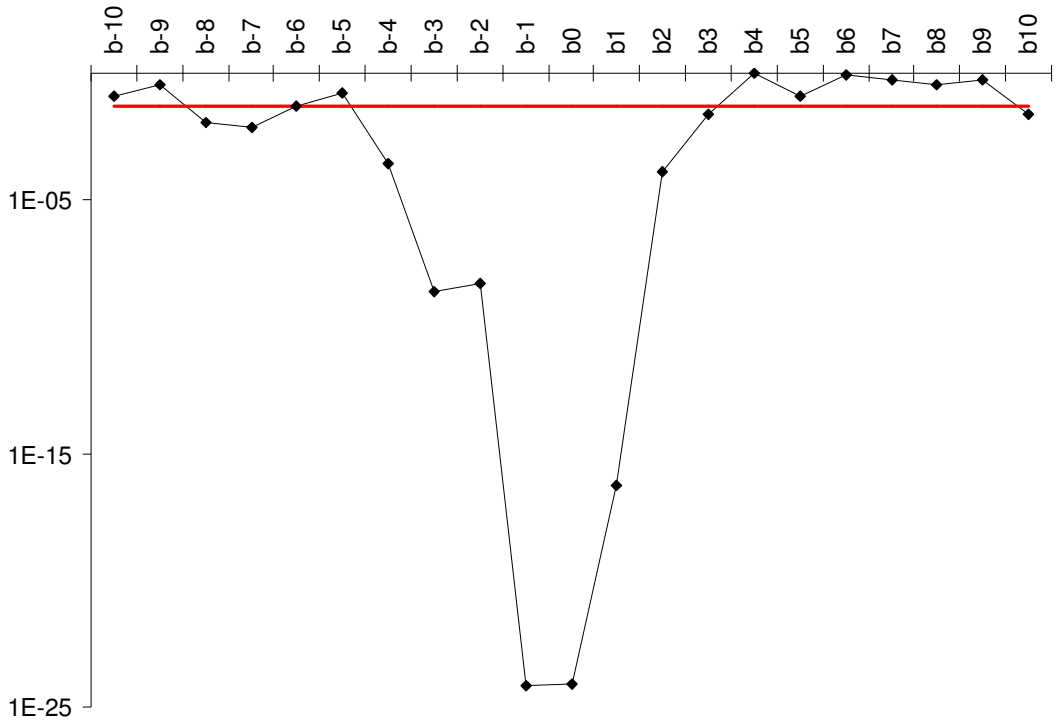


Figure 38 : Importance du contexte flanquant la base affectée par une substitution C>N.

Contrairement à la démarche utilisée dans l'étude préliminaire, les effectifs théoriques permettant d'effectuer un test du Chi² à chaque position ont été déterminés en fonction de la

composition moyenne des transcrits humains et de la distribution en phase des évènements de substitution.

La figure 38 montre que l'on retrouve une forte influence des 4 bases précédant et des 3 suivant la substitution. L'importante du contexte correspondant à la portion d'ADN ouverte par l'ARNpol II pendant la transcription est ainsi confirmée sur l'ensemble du transcriptome.

3.5.1.2 Contexte n-uplet des délétions

L'ensemble de données analysé correspond aux délétions C>N obtenues après application du test de proportions (n = 2.912). Aucun événement de RefSeq litigieuse ni de SNP délétion n'a été observé à l'intérieur de ce sous-ensemble de positions.

Les bases T et C sont très majoritairement affectées par les évènements de délétion.

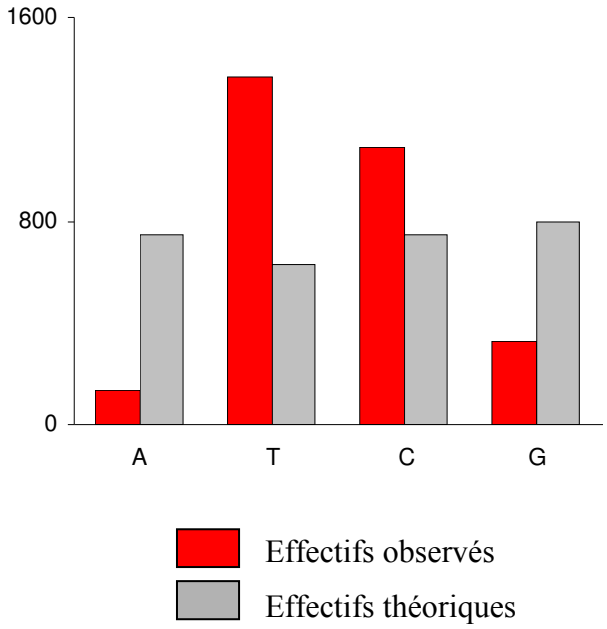


Figure 39 : Composition des bases affectées par les délétions C>N.

La nature des bases affectées n'est donc pas aléatoire ; la p-valeur du test de proportion entre les bases observées et les effectifs théoriques est inférieure à la précision machine.

90,8% des délétions surviennent au sein de répétitions de plus de deux bases de même nature. Il est intéressant de noter que les délétions affectant la base C surviennent essentiellement dans des doublets CC, alors que les délétions de T surviennent dans des doublets, triplets et quadruplets de T. Notons également que plus de 90% des délétions de T sont suivies de la base G. Enfin, lorsqu'une délétion touche la base G, il s'agit essentiellement de triplets GGG.

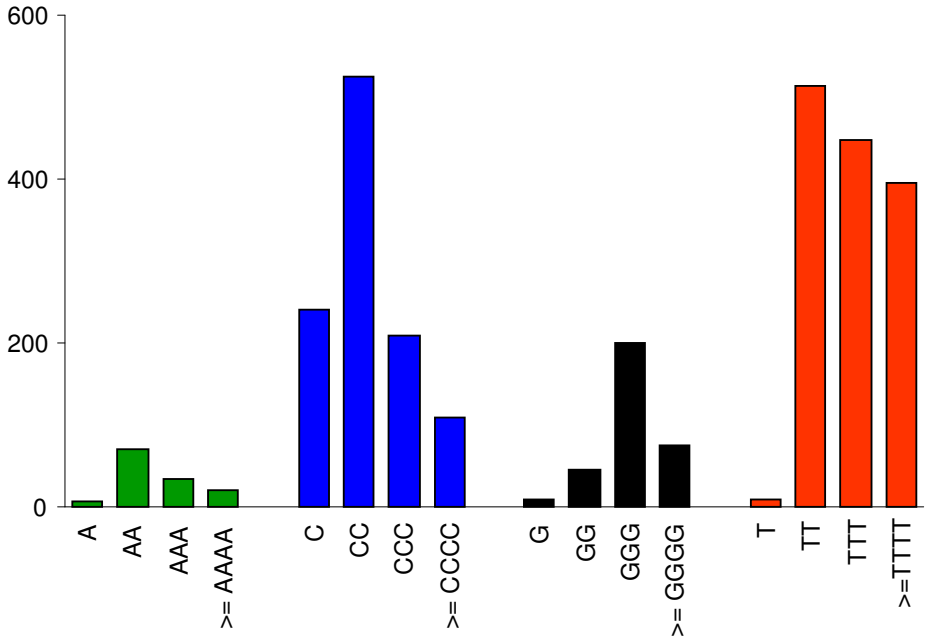


Figure 40 : Nature des répétitions impliquées dans les délétions C>N.

3.5.1.3 Contexte des insertions

Deux types d'insertions sont distingués dans cette étude, à savoir les insertions isolées et les insertions générant ou augmentant un n-uplet. Les secondes sont plus fréquemment rencontrées.

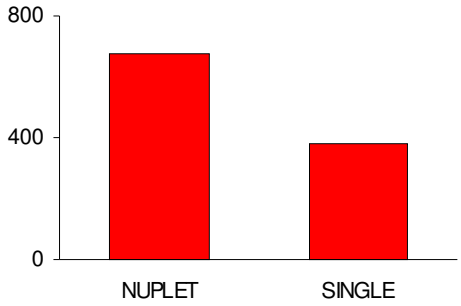


Figure 41 : Nombre de positions d'insertion C>N en fonction de la nature de l'insertion.

Comme pour les substitutions et les délétions, les positions dites litigieuses et les positions de SNP de même nature que l'événement d'infidélité sont éliminées. Dans le cas des insertions, très peu de positions sont éliminées (n=0 pour les insertions isolées et n=1 pour les insertions n-uplet).

D'autre part, seuls les évènements pour lesquels la base insérée représente plus de 75% de l'ensemble des bases insérées à cette position sont pris en considération.

La nature de la base insérée dans les cas d'événements isolés (n=268), pour l'ensemble des positions C>N, est la suivante :

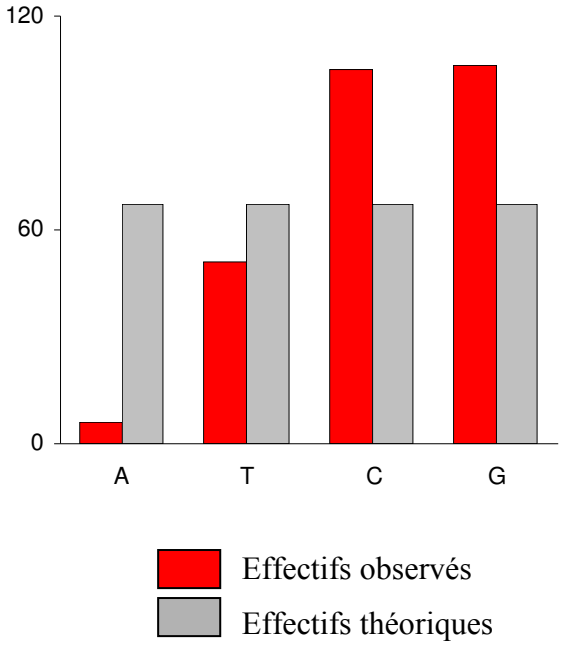


Figure 42 : Nature des bases insérées dans les cas d'insertion isolées C>N.

Par conséquent, la nature de la base insérée est différente de celle attendue par hasard (p-valeur = $3*10^{-22}$).

Il est intéressant de noter que, dans le cas d'évènements dits isolés, sur les 16 doublets de base entre lesquels une insertion peut survenir, seulement quatre sont représentés. Dans ces quatre cas, la base insérée n'est pas aléatoire :

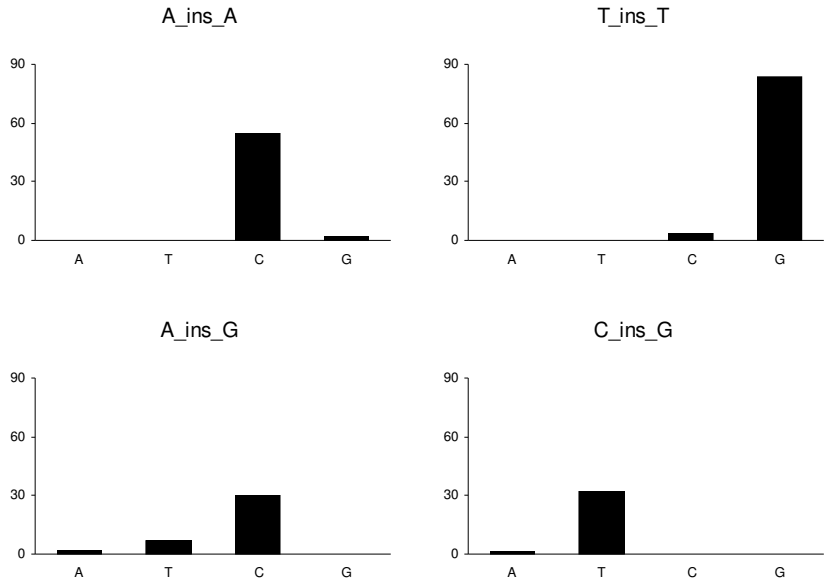


Figure 43 : Quatre types majoritaires d'insertions isolées.

Ainsi : AA → ACA, TT → TGT, AG → ACG et CG → CTG.

La nature de la base insérée dans les cas d'événements n-uplets (n=627), pour l'ensemble des positions C>N, est la suivante :

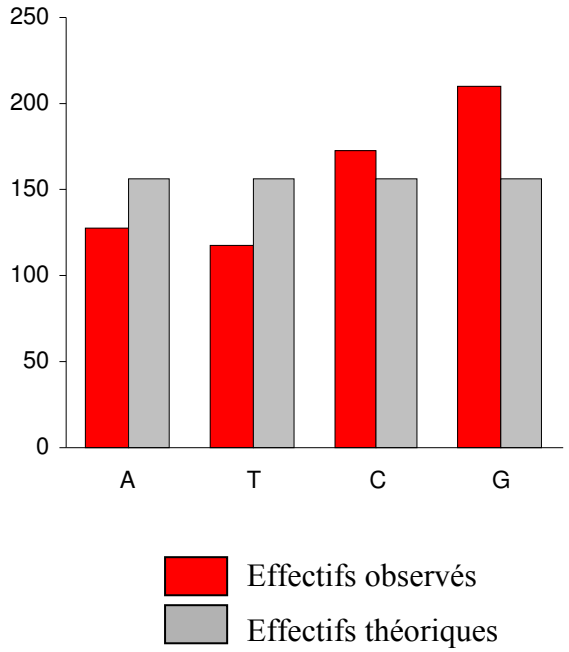


Figure 44 : Nature des bases insérées dans les cas d'insertion C>N survenant dans des répétitions de bases.

De la même manière que pour les insertions isolées, mais avec une intensité moindre, la nature de la base insérée est différente de celle attendue par hasard (p-valeur = $1*10^{-7}$).

Ainsi, les évènements d'infidélité de transcription surviennent de manière non aléatoire et sont soumis à différentes règles. Nous nous sommes alors demandé si les mutations somatiques survenaient, de la même manière, dans un contexte d'ADN spécifique.

3.5.2 Mutations somatiques

Une mutation somatique est une mutation qui affecte un gène d'une cellule somatique, par opposition à une mutation germinale, qui se produit au sein d'un gamète. Les mutations somatiques sont donc transmises à la descendance cellulaire mais ne sont pas héréditaires. Lorsque l'ADN d'une cellule somatique est modifié et qu'il en résulte une mutation, les

cellules issues de cette cellule mutée forment un clone de cellules mutées qui présentent des caractères différents de ceux du tissu d'origine.

Les mutations somatiques impliquées dans le cancer affectent les oncogènes, les gènes suppresseurs de tumeurs, ainsi qu'un grand nombre de gènes jusqu'à présent non connus pour leur implication dans la cancérogénèse [38,15].

3.5.2.1 Étude de tumeurs du sein et du colon

L'équipe de B. Vogelstein a réalisé une étude d'identification de mutations somatiques par séquençage à large échelle du génome de tumeurs du colon et du sein [15].

La démarche utilisée est la suivante :

- ✓ Choix de la séquence des amorces pour la PCR

La banque utilisée est la banque CCDS (consensus coding sequences) du NCBI (www.ncbi.nlm.nih.gov/CCDS) considérée par les auteurs comme étant le jeu de données de séquences codantes le plus fiable. Les exons codants de 14.661 transcrits, ainsi que les sites d'épissage (quatre bases introniques) ont été considérés. Plus de 100.000 séquences d'amorces ont été choisies afin d'amplifier 21 Mb de séquence génomique.

- ✓ PCR et séquençage

11 tumeurs mammaires, 11 tumeurs colorectales (lignées cellulaires ou xénogreffes) et deux échantillons non cancéreux ont été analysés. 3 millions de PCR ont été réalisées, permettant d'obtenir la séquence de 465 Mb de tumeur.

- ✓ Analyse des mutations

Les mutations silencieuses, les mutations retrouvées dans les échantillons non cancéreux, les SNPs connus, les mutations germinales et les mutations liées à des séquences homologues ont été filtrés. Par ailleurs, près de la moitié des mutations identifiées ont été éliminées après inspection visuelle des traces des chromatogrammes. Les deux tiers des mutations n'ayant pas été éliminées par l'ensemble de ces filtres ont été confirmées par reséquençage.

Cette phase de découverte a abouti à l'identification de 1.307 mutations localisées sur 1.149 gènes différents, et a été suivie d'une phase de validation portant sur 24 autres tumeurs. Cette seconde phase a impliqué 453.024 PCR supplémentaires concentrées sur les gènes identifiés par la phase de découverte et a permis l'identification de 365 mutations supplémentaires.

Ainsi, l'équipe de B. Vogelstein décrit 921 mutations somatiques de tumeurs du sein et 751 de tumeurs du colon.

Le taux de mutations somatiques est donc de 1.672 mutations pour 542 Mb de tumeurs séquencées, soit 3 mutations par Mb.

Une des conclusions de l'article est que les mutations somatiques correspondant à des transitions C : G → A : T sont plus fréquemment rencontrées dans les tumeurs du colon que dans les tumeurs du sein.

3.5.2.2 Cas particulier des transitions C : G → A : T

Les transitions C : G → A : T situées au niveau de sites 5'-CpG-3' sont reconnues comme étant les mutations somatiques les plus fréquentes dans le cancer colorectal [13]. Ce phénomène peut rendre compte d'une méthylation plus importante des îlots CpG ou être la conséquence d'une plus forte exposition des cellules du colon aux agents carcinogènes[140,141]. Dans le premier cas, une cytosine méthylée peut subir une désamination spontanée en thymine et échapper aux systèmes de contrôle [142,143]. Dans le second cas, des modifications chimiques de la guanine par différents agents alkylants génèrent une incorporation préférentielle de thymine et non de cytosine lors de la réplication [144-146]. L'insertion d'alanine en face de cette thymine erronée, soit par l'action d'enzymes réparatrices, soit par un nouveau cycle de réplication, génère des transitions C : G → A : T.

Les transitions C : G → A : T sont donc des phénomènes pour lesquels il existe des mécanismes identifiés, contrairement aux autres types d'évènements. Nous nous focaliserons donc sur les mutations autres que C : G → A : T. Par ailleurs, nous ne considérerons par la suite que les évènements exoniques.

3.5.2.3 Analyse du contexte d'ADN

L'objectif de ce paragraphe est d'étudier les bases flanquant les mutations somatiques afin de déterminer si elles ont un impact ou non sur la survenue du phénomène mutatoire.

Les mutations autres que C : G → A : T situées dans les exons sont au nombre de 783. Les auteurs ayant choisi de se concentrer sur les mutations non silencieuses, il existe un biais de phase important. En effet, les mutations ayant un impact codant sont préférentiellement situées en phases 1 et 2.

Nous avons commencé par définir une composition de référence dépendant de la phase considérée. L'ensemble des séquences CCDS du NCBI a été téléchargé dans les conditions décrites dans [15], *i.e.* à la date du 03/03/05.

Deux jeux de données sont construits :

- ✓ G1 : les compositions des trois phases sont déterminées sur un échantillon de positions tirées aléatoirement parmi l'ensemble des séquences CCDS.

- ✓ G2 : les compositions sont déterminées sur les séquences des transcrits présentant des mutations somatiques (après élimination des positions correspondant à ces mutations et de leurs 20 bases flanquantes).

	G1				G2		
	Phase 1	**Phase 2**	**Phase 3**		**Phase 1**	**Phase 2**	**Phase 3**
A	1102 (27.6%)	1236 (30.9%)	763 (19.1%)		15807 (26.3%)	18654 (31%)	11422 (19%)
T	672 (16.8%)	1029 (25.7%)	822 (20.6%)		10181 (17%)	15426 (25.7%)	13326 (22.2%)
C	962 (24%)	959 (24%)	1237 (30.9%)		14909 (24.8%)	14467 (24.1%)	18112 (30.1%)
G	1264 (31.6%)	776 (19.4%)	1178 (29.4%)		19189 (31.9%)	11555 (19.2%)	17226 (28.7%)

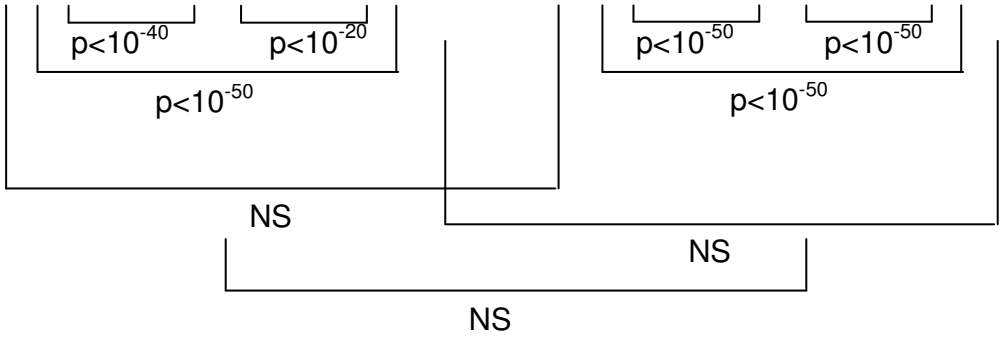


Table 20 : Composition théorique en bases des trois phases déterminée par deux approches.
La première approche G1 donne la composition de bases sélectionnées aléatoirement au sein d'exons codants, dont 4.000 positions en phase 1, 4.000 en phase 2 et 4.000 en phase 3. La seconde approche G2 donne la composition en bases des exons contenant les mutations somatiques suivant leur phase (à l'exception des 20 bases entourant la position exacte de la mutation).
Les p-valeurs obtenues après application de tests du Chi² sont données pour les deux approches. Les tests sont réalisés entre la composition des bases en phase 1 contre celle des bases en phase 2, puis la phase 1 contre la phase 3 et enfin la phase 2 contre la phase 3. Enfin, les p-valeurs des tests réalisés entre les deux approches sont données.

Les deux méthodes de détermination de composition théorique de bases selon la phase donnent des résultats identiques : la composition en bases est différente d'une phase à l'autre, il convient donc d'en tenir compte dans l'analyse.

La seconde partie du travail a consisté à définir précisément la distribution des 783 mutations et des bases voisines en termes de phase et de brin chromosomique.

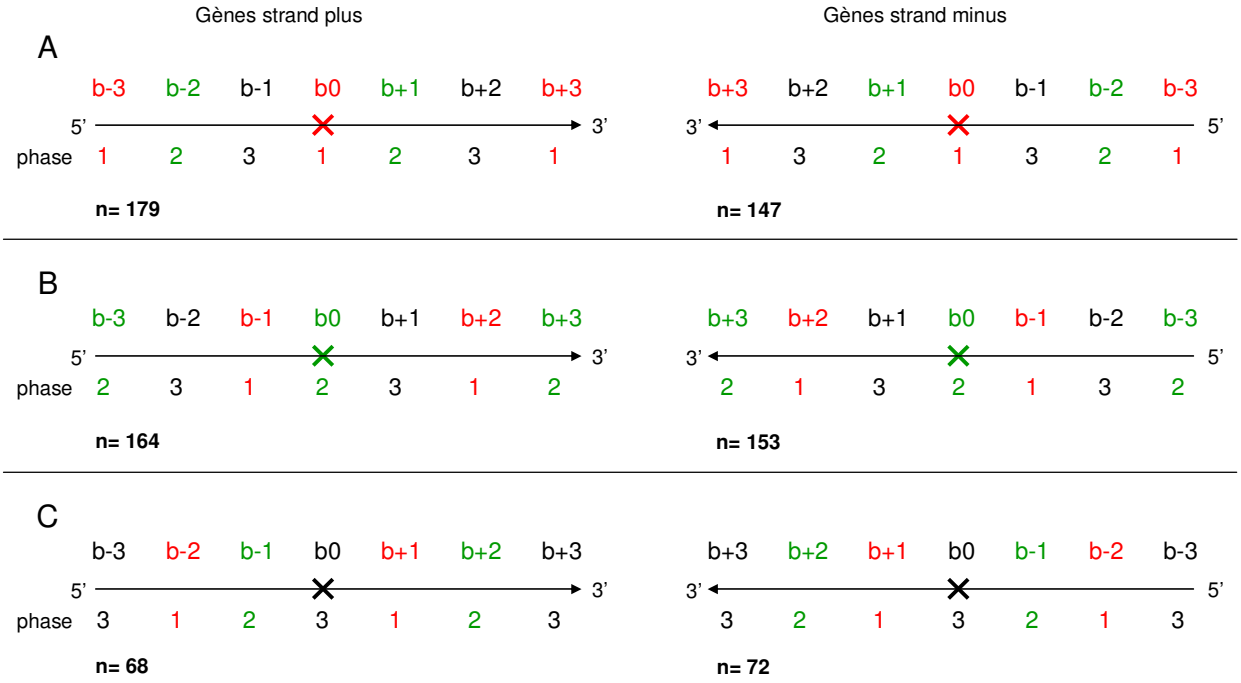


Figure 45 : Séparation des mutations somatiques autres que les transitions C : G → A : T en fonction de la phase et de l'orientation du gène.

Les mutations somatiques décrites sont extraites de Sjoblom et al. (Table S4). Les doublons, les mutations situées dans les introns, les mutations autres que les substitutions d'une base et les transitions G→A et C→T ont été éliminés. Les 783 mutations restantes ont été divisées en 6 groupes : A. les mutations portant sur la première base d'un codon et localisées sur un gène situé sur le brin plus ou minus du chromosome. B. Les mutations portant sur la deuxième base du codon et C. sur la troisième base.

b0 représente la base subissant la mutation sur l'ARNm et b-3 à b+3 sont les bases situées en amont et en aval de la mutation. La représentation tient compte de l'orientation du gène sur le chromosome.

52% des mutations étudiées sont localisées dans un gène situé sur le brin *plus* d'un chromosome et 48% sur le brin *moins*. Les mutations affectent donc les deux brins de la même manière (test du Chi² non significatif). 42 et 40% des mutations sont situées en phases 1 et 2 respectivement et seulement 18% en phase 3 ; le biais est donc réel (Chi², $p<10^{-20}$) et dû à la méthodologie de sélection de mutations somatiques ayant un impact codant.

Il est important de constater que les proportions de mutations situées en phase 1, 2 et 3 sont les mêmes quel que soit le brin chromosomique considéré. La phase des bases suivant ou précédant la mutation n'étant pas liée au brin chromosomique, il est possible de regrouper les mutations situées sur les deux brins, à partir du moment où l'on prend en compte le sens de lecture du gène. La figure suivante montre qu'il n'y a pas de différence de contexte entre les mutations autres que C : G → A : T suivant le brin chromosomique considéré.

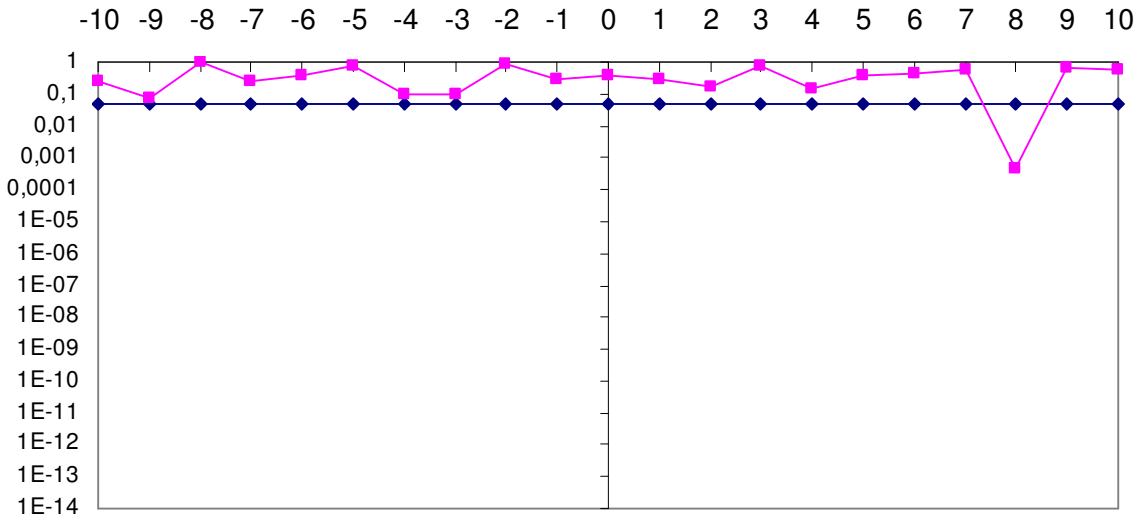
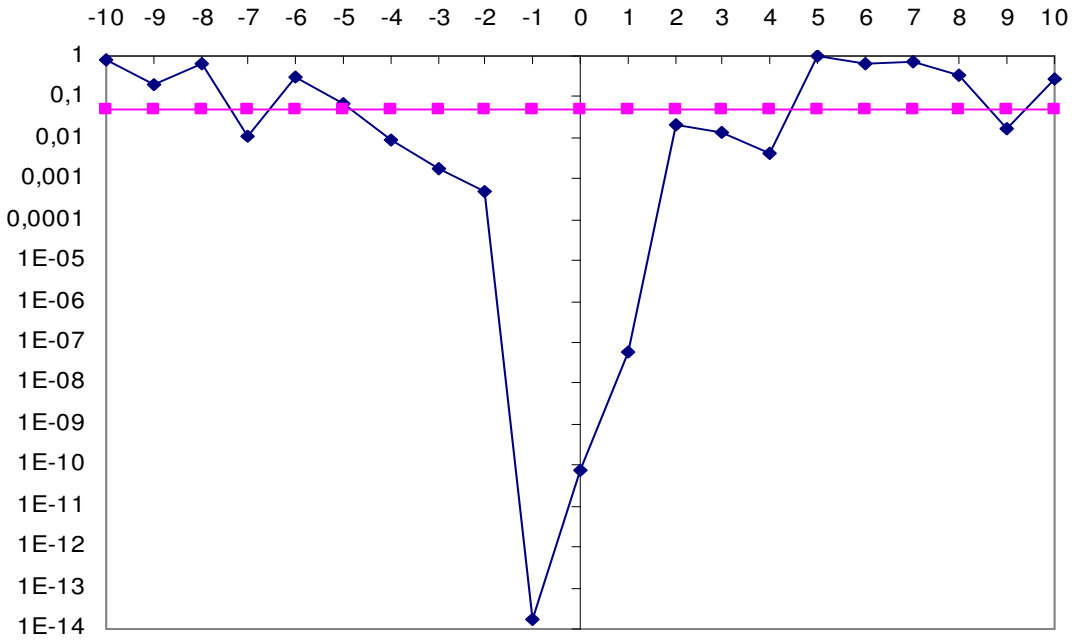


Figure 46 : Différence du contexte flanquant les mutations somatiques autres que les transitions C : G → A : T suivant l'orientation du gène sur le chromosome.

Les p-valeurs du test du Chi² réalisé entre la composition en bases des mutations portées par des gènes situées sur le brin plus (n=411) et celle des mutations portées par des gènes situés sur le brin minus (n=372) sont représentées. La ligne bleue représentent le seuil de significativité (p=0,05).

Nous avons ensuite cherché à savoir si le contexte d'ADN flanquant les mutations somatiques était différent ou non du contexte d'ADN de positions non affectées par des mutations somatiques.

L'analyse a d'abord été réalisée sur les mutations somatiques autres que les transitions C : G → A : T. La figure suivante montre que le contexte lié à la survenue de ces évènements s'étend sur 9 bases, incluant les 4 bases précédant et les 4 bases suivant la mutation.



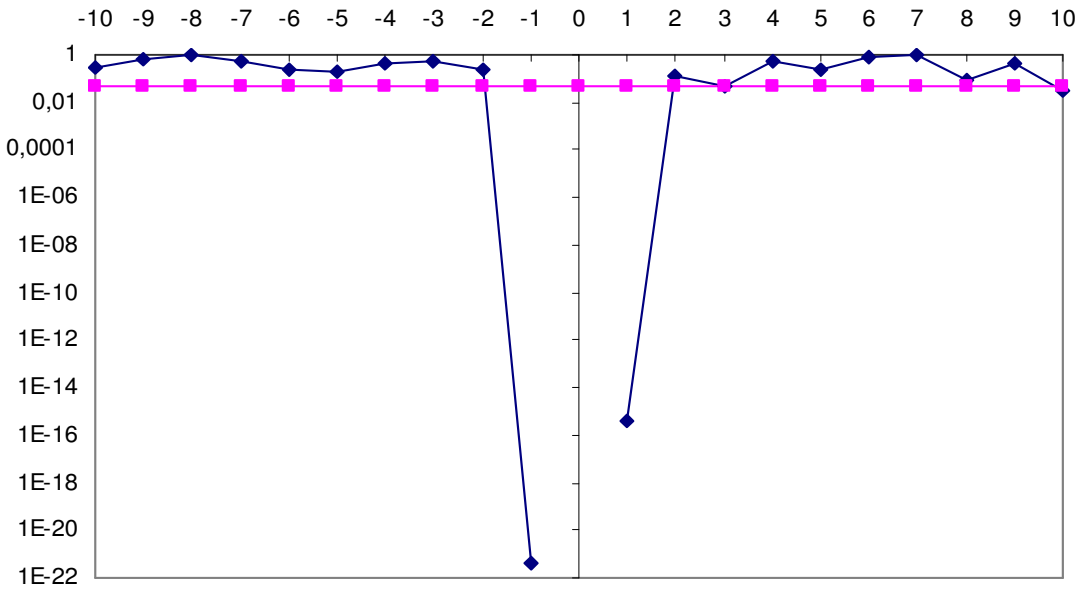
	b-4		b-3		b-2		b-1		b0		b+1		b+2		b+3		b+4	
	NM	M	NM	M	NM	M	NM	M	NM	M	NM	M	NM	M	NM	M	NM	M
A	24	24	27	33	25	29	24	34	27	24	25	34	24	28	27	30	25	27
T	21	23	21	21	23	25	21	24	21	13	23	24	21	20	21	23	23	27
C	27	22	25	21	27	26	27	18	25	30	27	22	27	23	25	21	27	22
G	28	30	26	25	25	20	28	24	26	34	25	20	28	29	26	26	25	25

Figure 47 : Importance du contexte flanquant les mutations somatiques autres que les transitions C : G → A : T.

p-valeurs des tests du Chi² entre la composition des mutations somatiques autres que les transitions C : G → A : T (M) et la composition théorique attendue (NM). La ligne rose représentent le seuil de significativité (p=0,05). Les compositions de bases présentant une différence significative sont données dans le tableau. Les bases plus fréquentes sont en jaune et les moins fréquentes en bleu.

Les bases les plus fréquemment affectées restent, après élimination des transitions C : G → A : T, les bases C et G. Les bases flanquant les mutations sont préférentiellement des adénines.

Le contexte des mutations somatiques correspondant à des transitions C : G → A : T a alors été comparé au contexte de positions G1 décrites précédemment. Plus précisément, nous avons construit un tableau de composition de bases théorique, tenant compte à la fois du nombre de mutations somatiques en phase 1, 2 et 3 et de la composition théorique de ces 3 phases. Il apparaît très nettement que la base précédant et la base suivant les mutations somatiques ne sont pas aléatoires. Cela s'explique aisément par le fait que les transitions C : G → A : T surviennent très préférentiellement au sein d'îlots CpG [147].



	b-1		b0		b+1	
	NM	M	NM	M	NM	M
A	22	12	28	0	26	24
T	20	21	21	0	24	16
C	28	44	25	42	27	23
G	30	23	26	58	23	37

Figure 48 : Importance du contexte flanquant les mutations somatiques correspondant à des transitions C : G → A : T.

Même légende que celle de la figure 47.

Ces données indiquent que l'on peut définir deux types de mutations somatiques :

- ✓ les mutations somatiques correspondant à des transitions C : G → A : T, mutations qui sont liées à un contexte d'ADN très limité et qui sont expliquées par des altérations chimiques de C et G,
- ✓ les mutations somatiques autres que les transitions C : G → A : T, qui sont conditionnées par un large contexte d'ADN correspondant à la portion d'ADN ouverte au cours de la transcription.

PUBLICATION 2 :

Colon and breast cancer somatic mutations other than C:G - T:A transitions are non random events conditioned by DNA context.

Brulliard Marie, Moncuquet Philippe, Collignon Olivier, Jacquenet Sandrine, Ogier Virginie, Roitel Olivier, Thouvenot Benoît, Bihain Bernard E.

En cours de rédaction.

L'étude bioinformatique portant sur l'ensemble des données ESTs humaines disponibles sur le NCBI confirme l'augmentation de l'hétérogénéité des transcrits issus de tissus cancéreux par rapport aux transcrits issus de tissus sains. L'étude à grande échelle a permis en outre de mettre au jour les évènements de délétion et d'insertion. L'observation bioinformatique la plus marquante porte sur les délétions d'une base obtenues par application du test de proportions après application du filtre -10/+10. En effet, les positions C>N, *i.e.* présentant significativement plus de délétions dans les ESTs cancéreuses que dans les ESTs normales, sont 14 fois plus nombreuses que les positions N>C. Cet ensemble de positions C>N a ainsi été retenu pour valider biologiquement l'hypothèse d'infidélité de transcription.

L'analyse bioinformatique a été décrite dans les parties 2 et 3 de ce manuscrit. Les résultats sont intéressants mais restent des hypothèses tant qu'ils n'ont pas été validés biologiquement. Les parties 4 et 5 vont s'attacher à établir différentes preuves de concept biologiques du phénomène bioinformatique d'infidélité de transcription. Il me semble important de rappeler que les résultats présentés dans ce manuscrit sont les résultats de toute une équipe. Les prédictions bioinformatiques sont réalisées par le groupe de bioinformatique et biostatistiques et les expériences biologiques sont réalisées par les différentes équipes de biologistes de Genclis. Il est nécessaire de présenter les preuves de concept biologiques car d'une part elles donnent plus de poids aux résultats bioinformatiques, et d'autre part, elles permettent d'apporter des perspectives nouvelles au concept d'infidélité de transcription.

4 Prédictions bioinformatiques et résultats biologiques

4.1 Détection d'un ARNm présentant une délétion à une position prédite par la bioinformatique

4.1.1 Prédiction bioinformatique

La délétion située en position 447 du transcrit NM_005507.2 du gène de la cofiline CFL1, soit dans un triplet TTT suivi de G (*i.e.* le contexte le plus commun), a été choisie.

GACGACCCCTACGCCACCTTTGTCAAGATGCTGCCAGATAA

Le tableau de contingence de cette délétion est le suivant :

	T	délétion
ESTs normales	622	3
ESTs cancéreuses	1167	57

P = 8,1 E-7

Table 21 : Tableau de contingence de la délétion considérée sur le transcrit de la cofiline.

4.1.2 Preuve de concept

- ✓ Préparation du plasmide

Le plasmide pBAD (Invitrogen) a été utilisé de manière à avoir un promoteur inductible (promoteur Ara BAD) placé en amont de la séquence clonée. La séquence du peptide alpha amplifiée à partir de plasmide pBS-SK+ a été introduite dans le plasmide pBAD pour former le plasmide alpha-pBAD. Cette séquence n'est pas clonée dans la même phase que l'ATG présent dans le site de clonage. Ainsi, par construction, si aucune séquence étrangère n'est introduite entre les sites de clonage Nhe I et Nco I, le peptide alpha n'est pas exprimé et, par conséquent, la colonie *E. coli* renfermant la construction est blanche.

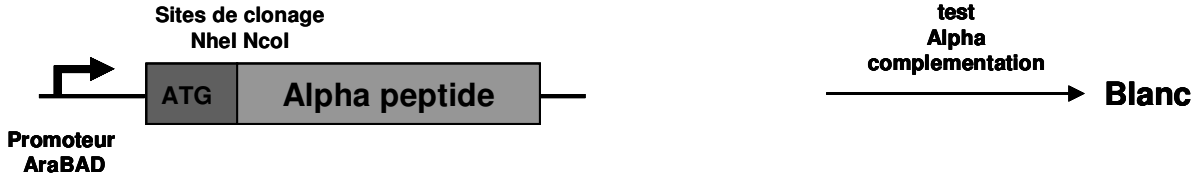


Figure 49 : Représentation de la construction utilisée.

✓ Clonage

Les ADNc étudiés proviennent d'une part d'un individu sain à qui l'on a prélevé du tissu pulmonaire et d'autre part d'un individu atteint de cancer du poumon à qui l'on a prélevé un fragment de tumeur pulmonaire ainsi qu'un fragment de poumon sain adjacent (Biochain Inc.). Les ADN génomiques ont été séquencés et ne présentent pas de différence par rapport à la séquence de référence :

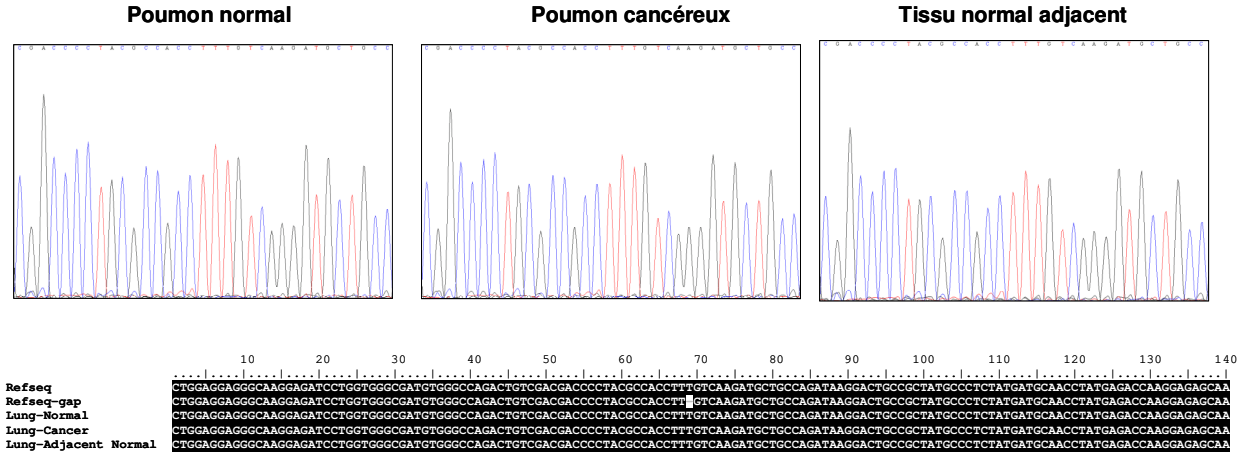


Figure 50 : Séquençage de l'ADN génomique.

Les ADNc ont été amplifiés par PCR en utilisant des amorces spécifiques du gène CFL1 et une polymérase hautement fidèle (Phusion de Finnzyme). Les ADNc ont ensuite été purifiés et digérés par les enzymes de restriction NcoI et NheI (Biolabs). Les fragments résultants ont alors été ligaturés dans le plasmide alpha-pBAD, lui-même ayant été digéré par les mêmes enzymes.

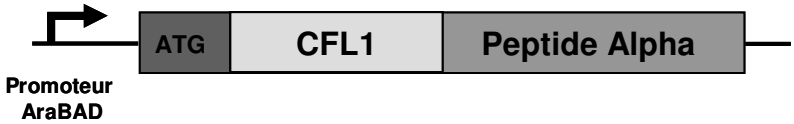


Figure 51 : Intégration du fragment de CFL1 dans le plasmide.

✓ Screening des colonies

Lorsqu'un fragment CFL1 normal est cloné, le peptide alpha n'est pas en phase avec l'ATG et la colonie est blanche. Si le fragment CFL1 cloné contient une délétion, la phase est restaurée. Une protéine de fusion CFL1-peptide alpha est alors produite, la beta-galactosidase est activée et la colonie est bleue.

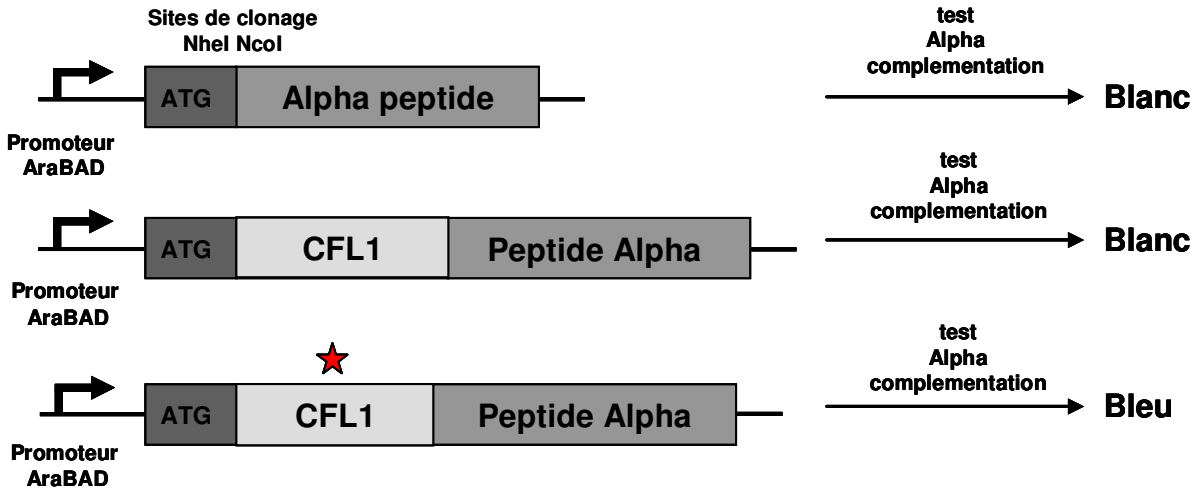


Figure 52 : Représentation de la construction utilisée après clonage du fragment de CFL1 portant ou non la délétion d'intérêt.

L'insert du plasmide des clones bleus a alors été séquencé (GATC Biotech).

Le résultat du séquençage est aligné à la séquence de référence de la cofiline. Un des clones obtenus présentait la délétion prédite par l'étude bioinformatique.

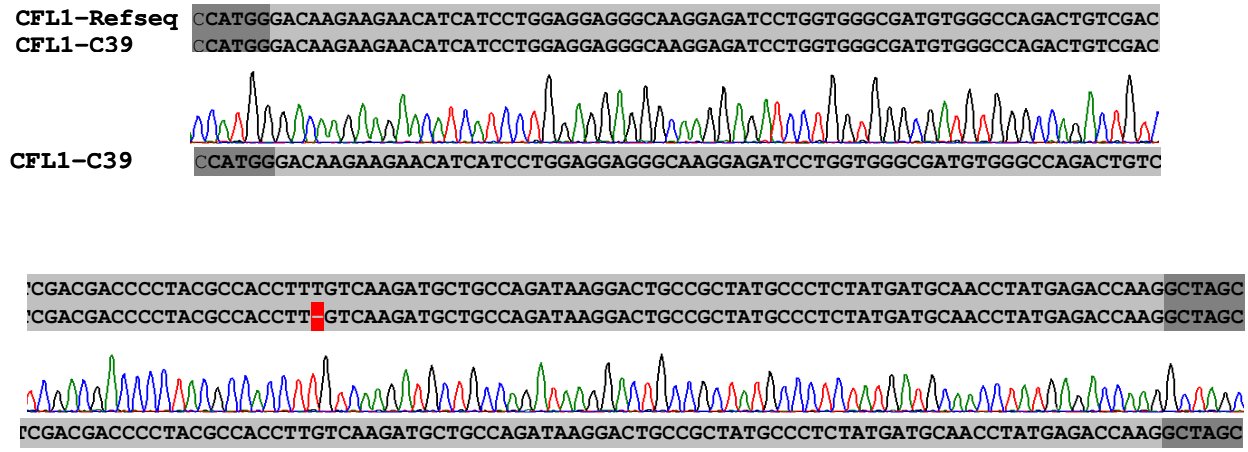


Figure 53 : Séquençage du clone porteur de la délétion prédite.

Des expériences similaires réalisées sur du tissu pulmonaire d'un individu non cancéreux et sur un tissu pulmonaire adjacent à la tumeur n'ont pas permis de mettre en évidence la présence d'ARNm erronés.

4.1.3 Conclusion

Il a donc été possible d'identifier, au sein d'un tissu pulmonaire cancéreux, un ARNm portant une délétion à la position exacte prédite par l'analyse bioinformatique, *i.e.* la position 447 du transcrit de CFL1.

L'existence d'ARNm humains porteurs de délétions d'une base à une position C>N prédite par la bioinformatique est donc validée. La question qui se pose alors, si l'on considère l'existence d'ARNm infidèles, est de savoir si ces ARN seront traduits et par conséquent, si l'on peut identifier des protéines issues d'ARN infidèles.

4.2 Infidélité de transcription affectant le codon stop

4.2.1 Définition et nature des stops alternatifs

Les séquences de référence ARNm (n=33.121) ont été téléchargées sur le site du NCBI au format fasta, ainsi que les fiches Genbank correspondantes. Ces fiches contiennent les positions des codons *start* et *stop* et permettent donc de localiser le stop naturel de chacune des séquences. La séquence des 33.121 codons stop a été ainsi déterminée (*i.e.* TGA, TAA ou TAG). Ensuite, l'UTR 3' a été balayée dans le but d'identifier la nature des codons stop dits alternatifs, *i.e.* situés en aval du stop naturel et en phase avec ce dernier. La table ci-dessous montre l'usage des 3 codons stop pour le stop naturel d'une part et pour l'ensemble des stops alternatifs d'autre part :

		nb stop	TAA	TAG	TGA	TAA	TAG	TGA
stop naturel	**0**	33121	9295	7604	16222	28%	23%	49%
stop alternatifs	**1**	26268	8171	5636	12460	31%	21%	47%
	2	24368	8361	5375	10632	34%	22%	44%
	3	22521	8406	4979	9136	37%	22%	41%
	4	20857	7750	4720	8387	37%	23%	40%
	5	19391	7082	4377	7932	37%	23%	41%
	6	18131	6774	4230	7127	37%	23%	39%
	7	16910	6425	3760	6725	38%	22%	40%
	8	15866	6018	3613	6235	38%	23%	39%
	9	14909	5694	3560	5655	38%	24%	38%
	10	14084	5520	3213	5351	39%	23%	38%

Table 22 : Usage des trois codons stop chez l'homme au niveau du codon stop naturel et des codons stop situés en phase dans l'UTR3'.

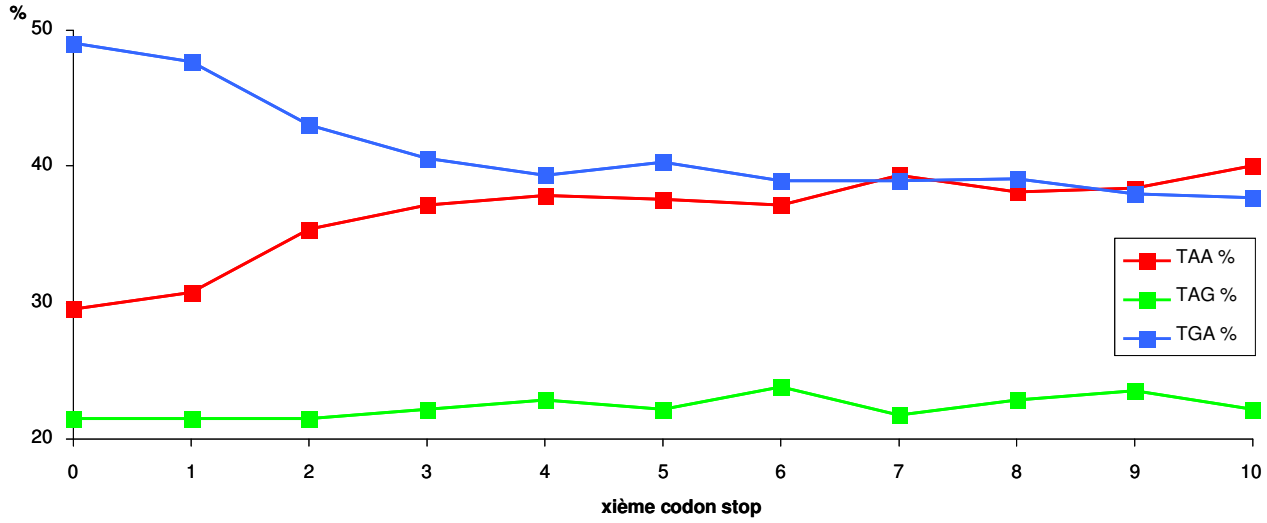


Figure 54 : Représentation de l'usage des trois codons stop chez l'homme.

Remarque : les résultats sont identiques en considérant comme sous-ensemble de départ de l'analyse les 11.000 séquences ARNm «reviewed» par NCBI, *i.e.* les séquences présentant le degré de fiabilité le plus haut.

Ainsi, il existe un biais d'usage du code génétique au niveau du codon stop naturel : TGA est largement majoritaire. Il est particulièrement intéressant de noter que ce biais existe encore pour le premier stop alternatif et dans une moindre mesure pour le deuxième. Il existe donc une pression évolutive portant sur le codon stop naturel mais aussi sur le premier, voire le deuxième stop situé dans l'UTR 3' et en phase avec le stop naturel.
Ces résultats semblent indiquer que l'évolution tend à avantager une extension de la traduction au delà du stop naturel.

Ce résultat peut être lié à l'infidélité de transcription de manière intéressante. En effet, l'analyse préliminaire portant sur 17 gènes abondamment exprimés a montré que 9 transcrits sur 17 présentaient une substitution C>N au niveau du codon stop. Les ARNm portant une substitution dans le codon stop seront donc traduits jusqu'au premier stop alternatif rencontré en phase. Le peptide supplémentaire ainsi traduit sera nommé Post-Stop Peptide ou PSP.

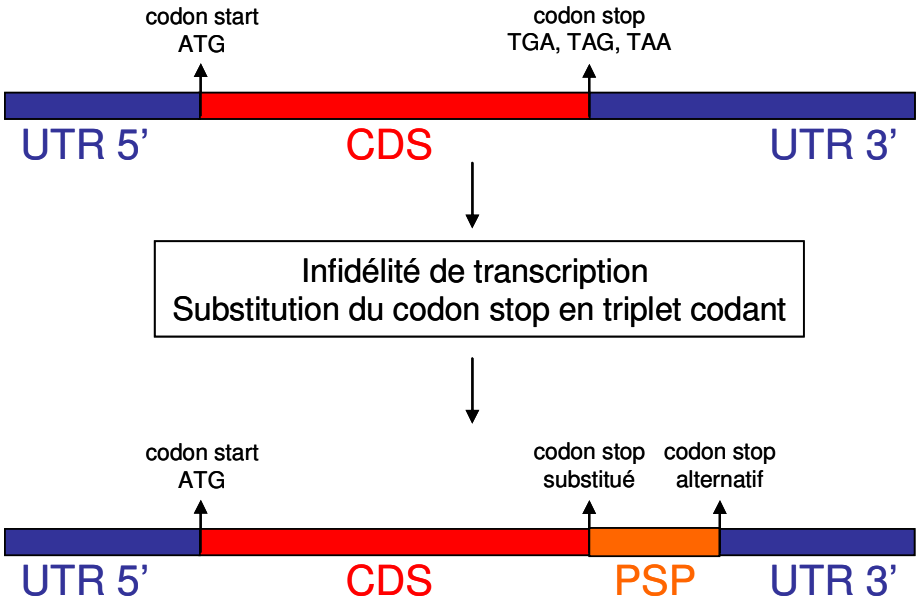


Figure 55 : Représentation schématique d'un PSP.

4.2.2 Preuve de concept

L'hypothèse d'extension de la traduction au-delà du stop naturel a été testée sur une protéine humaine, l'apolipoprotéine AII (ApoAII). L'ApoAII a été choisie pour les raisons suivantes :

- ✓ Protéine abondante du plasma et « *facilement* » purifiable,
- ✓ Pas de pseudogène connu,
- ✓ Pas de correspondance entre le peptide prédit et une traduction complète dans les 6 phases du génome humain,
- ✓ Stop canonique le plus commun, *i.e.* TGA.



Figure 56 : Définition du PSP de l'ApoAII humaine.

Des anticorps, dirigés d'une part contre la partie N-terminale d'ApoAII et d'autre part contre le PSP, ont été produits et ont permis de mettre en évidence la présence d'ApoAII normale et

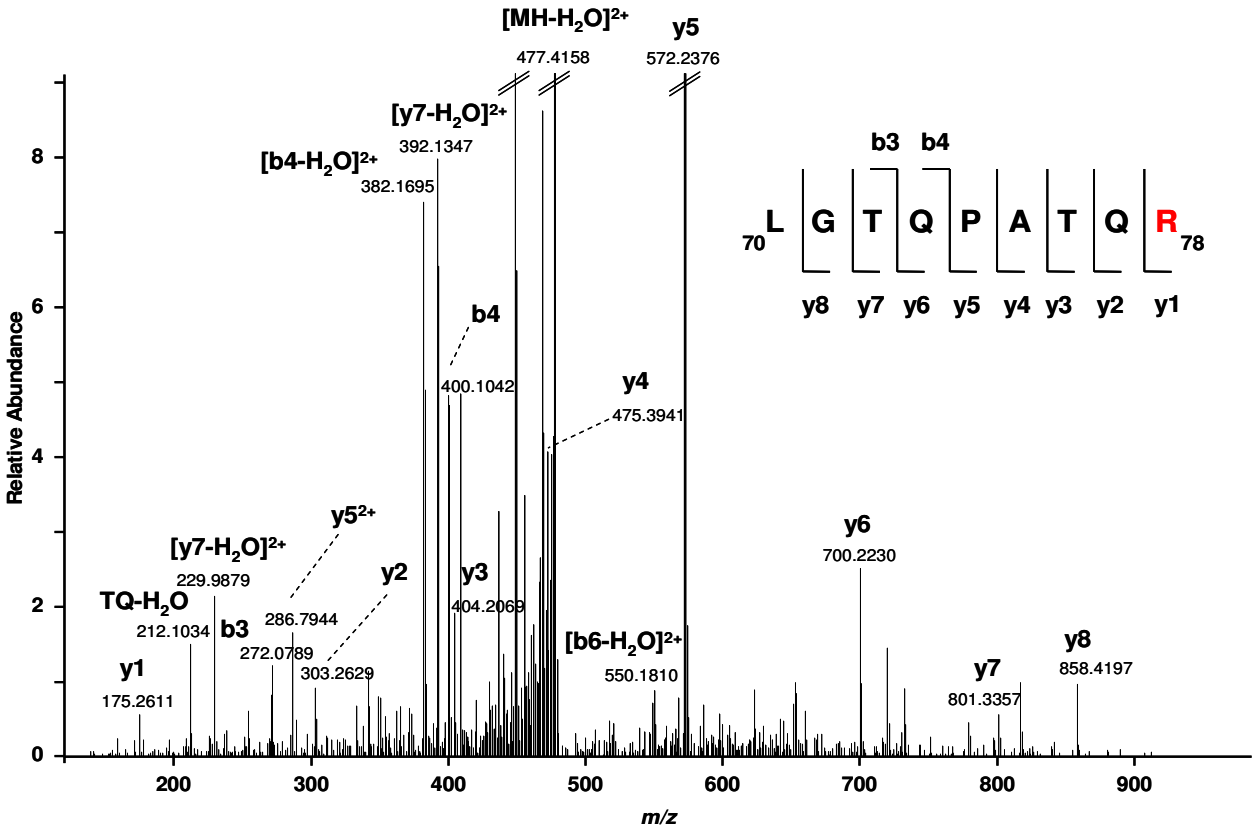
de l'isoforme étendue en carboxy-terminal (ApoAII-PSP) chez l'ensemble des 16 individus sains testés. La purification de cette isoforme a permis de confirmer l'existence du PSP par spectrométrie de masse en tandem (MS/MS). Cette confirmation a été réalisée par l'équipe de Bernard Monsarrat au sein de l'Institut de Pharmacologie et de Biologie Structurale de Toulouse.

$_{1}$QAKEPCVESLVSQYFQTVTDYGKDLMEKVKSPELQAEAKSYFEKSKEQLTPLIKK

AGTELVNFLSYFVELGTQPATQRSVQTIVFQPQLASRTPTGQS$_{98}$

Figure 57 : Séquence de l'ApoAII-PSP.

Les fragments de séquence soulignés correspondent aux peptides identifiés en MS/MS. Le résidu d'arginine (rouge) remplace le codon stop. Les lignes rouges correspondent aux peptides obtenus après digestion trypsique de l'ApoAII-PSP, les lignes vertes à ceux obtenus après double digestion trypsine / endoprotéase Glu-C (V8).



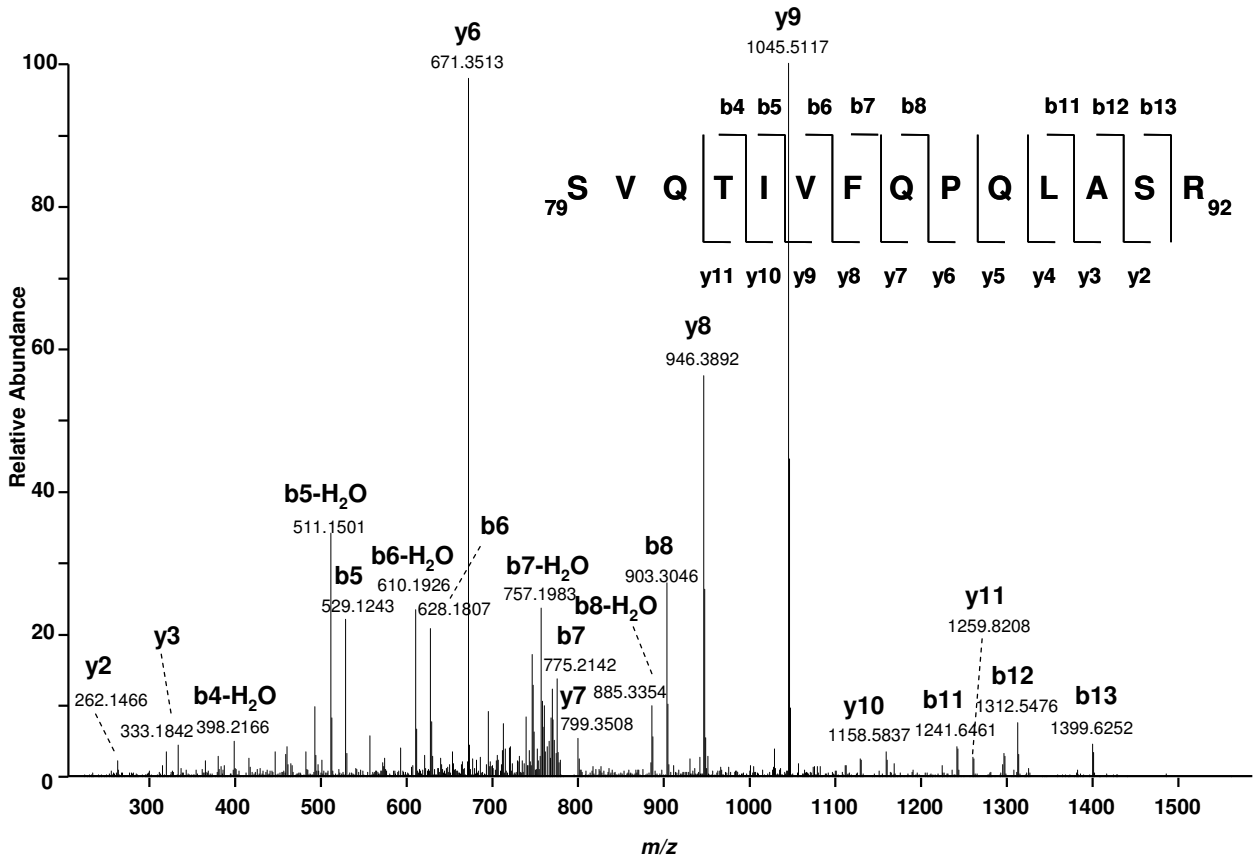


Figure 58 : Analyse des spectres de MS/MS après double digestion.

Les deux séquences peptidiques obtenues sont indiquées en haut à gauche de chacun des deux spectres. Le premier spectre donne la partie C-terminale de l'ApoAII, où un résidu d'arginine remplace le codon stop. Le deuxième donne une partie de la séquence du PSP.

La dérivation du codon stop existe donc chez les sujets normaux et est capable de générer une diversité protéique jusqu'alors insoupçonnée.

La preuve de concept portant sur la présence, dans le plasma humain, de protéines variantes issues d'ARN présentant une substitution dans le codon stop générant un nouveau cadre de lecture, est ainsi réalisée. On ne peut pas à ce stade exclure l'hypothèse d'infidélité de traduction, bien que l'on privilégie, sur la base de l'étude bioinformatique, la thèse d'infidélité de transcription.

Nous nous sommes ensuite intéressés aux conséquences des évènements d'infidélité de transcription les plus pertinents d'un point de vue bioinformatique, à savoir les délétions.

4.3 Conséquences d'une délétion sur la protéine et prédiction de peptides issus d'infidélité de transcription

4.3.1 Impact codant d'une délétion

Une délétion située dans la partie codante d'un ARNm provoque un décalage du cadre de lecture. La traduction de cet ARNm infidèle conduit ainsi à la production de protéines dont la séquence carboxi-terminale diverge de la séquence de référence dite normale.

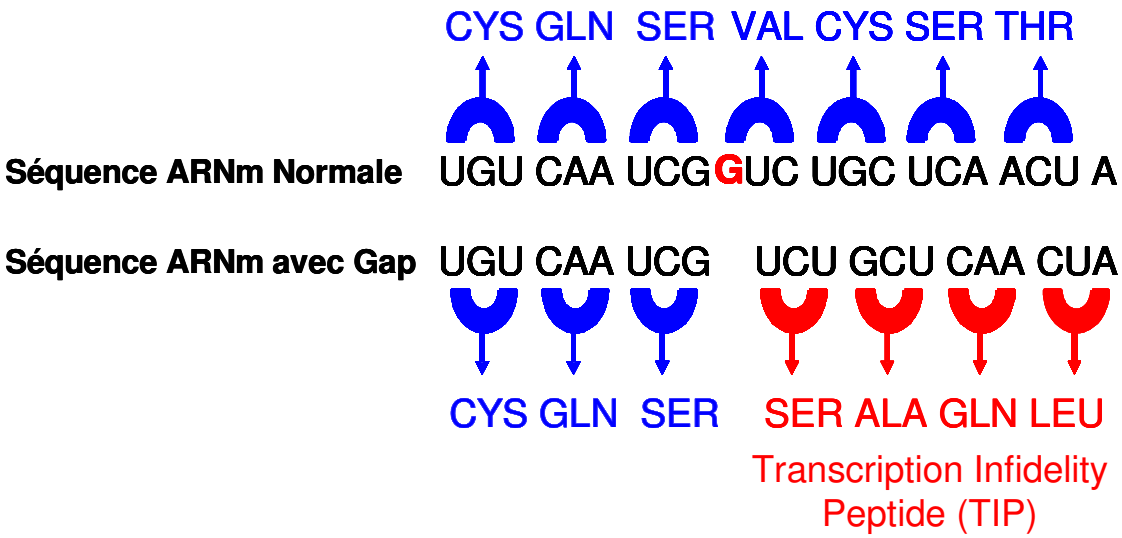


Figure 59 : Conséquence d'une délétion dans l'ARN sur la séquence protéique.

Les protéines aberrantes sont constituées d'une partie normale (en amont de la délétion) et d'une partie appelée Transcription Infidelity Peptide ou TIP (en aval de la délétion). La séquence des TIPs est définie par rapport à la séquence protéique de référence comme étant strictement différente de la protéine normale.

En dehors du cadre de lecture canonique, les codons stop sont plus fréquents. L'ARNm infidèle est donc traduit jusqu'au premier codon stop alternatif rencontré dans cette phase nouvellement définie.

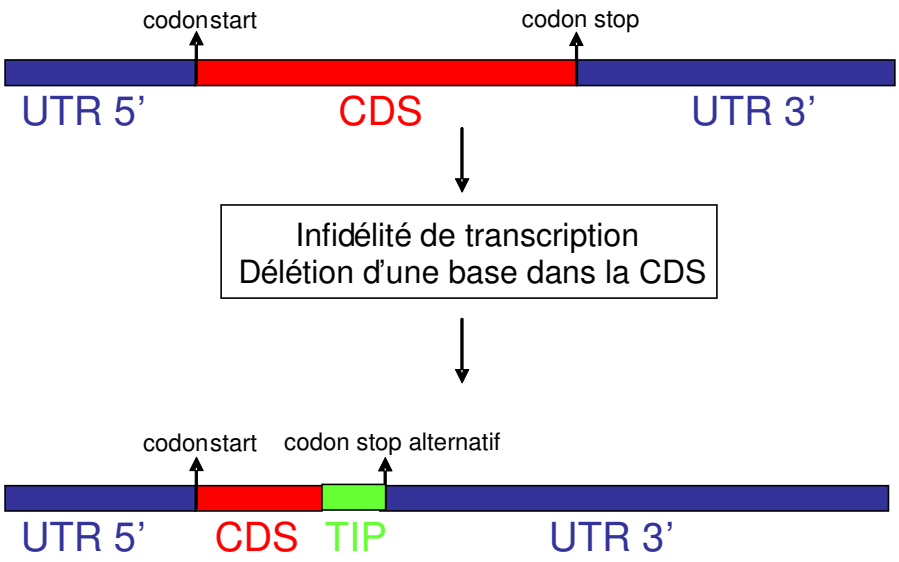


Figure 60 : Représentation schématique d'un TIP.

Remarque : les insertions ont un impact codant très semblable à celui des délétions. Néanmoins, nous nous intéresserons dans ce manuscrit aux délétions uniquement.

L'exemple du gap survenant sur CFL1 et présenté dans le paragraphe 4.1. conduit à la prédiction du TIP suivante :

ARNm de référence

GGCCGGCGGGAAGACTCCGTTACCCAGCGAGCGAGGCGGCGGCGCAGGGCCAGCGGACTCCATTTCCCGTCGGCT
CGCGGTGGGAGCGCCGGAAGCCCGCCCCACCCCTCATTGTGCGGCTCCTACTAAACGGAAGGGGCCGGGAGAGGC
CGCGTTCAGTCGGGTCCCGGCAGCGGCTGCAGCGCTCTCGTCTTCTGCGGCTCTCGGTGCCCTCTCCTTTTCGTT
TCCGGAAACATGGCCTCCGGTGTGGCTGTCTCTGATGGTGTCATCAAGGTGTTCAACGACATGAAGGTGCGTAAG
TCTTCAACGCCAGAGGAGGTGAAGAAGCGCAAGAAGGCGGTGCTCTTCTGCCTGAGTGAGGACAAGAAGAACATC
ATCCTGGAGGAGGGCAAGGAGATCCTGGTGGGCGATGTGGGCCAGACTGTCGACGACCCCTACGCCACCTTTGTC
AAGATGCTGCCAGATAAGGACTGCCGCTATGCCCTCTATGATGCAACCTATGAGACCAAGGAGAGCAAGAAGGAG
GATCTGGTGTTTATCTTCTGGGCCCCCGAGTCTGCGCCCCTTAAGAGCAAAATGATTTATGCCAGCTCCAAGGAC
GCCATCAAGAAGAAGCTGACAGGGATCAAGCATGAATTGCAAGCAAACTGCTACGAGGAGGTCAAGGACCGCTGC
ACCCTGGCAGAGAAGCTGGGGGGCAGTGCCGTCATCTCCCTGGAGGGCAAGCCTTTGTGAGCCCCTTCTGGCCCC
CTGCCTGGAGCATCTGGCAGCCCCACACCTGCCCTTGGGGGTTGCAGGCTGCCCCCTTCCTGCCAGACCGGAGGG
GCTGGGGGGATCCCAGCAGGGGGAGGGCAATCCCTTCACCCCAGTTGCCAAACAGACCCCCCACCCCCTGGATTT
TCCTTCTCCCTCCATCCCTTGACGGTTCTGGCCTTCCCAAACTGCTTTTGATCTTTTGATTCCTCTTGGGCTGAA
GCAGACCAAGTTCCCCCCAGGCACCCCAGTTGTGGGGGAGCCTGTATTTTTTTTAACAACATCCCCATTCCCCAC
CTGGTCCTCCCCCTTCCCATGCTGCCAACTTCTAACCGCAATAGTGACTCTGTGCTTGTCTGTTTAGTTCTGTGT
ATAAATGGAATGTTGTGGAGATGACCCCTCCCTGTGCCGGCTGGTTCCTCTCCCTTTTCCCCTGGTCACGGCTAC
TCATGGAAGCAGGACCAGTAAGGGACCTTCGATTAAAAAAAAAAAAGACAATAATAAAAA

Protéine de référence

MASGVAVSDGVIKVFNDMKVRKSSTPEEVKKRKKAVLFCLSEDKKNIILEEGKEILVGDVGQTVDDPYATFVKML
PDKDCRYALYDATYETKESKKEDLVFIFWAPESAPLKSKMIYASSKDAIKKKLTGIKHELQANCYEEVKDRCTLA
EKLGGSAVISLEGKPL

ARNm infidèle

GGCCGGCGGGAAGACTCCGTTACCCAGCGAGCGAGGCGGCGGCGCAGGGCCAGCGGACTCCATTTCCCGTCGGCT
CGCGGTGGGAGCGCCGGAAGCCCGCCCCACCCCTCATTGTGCGGCTCCTACTAAACGGAAGGGGCCGGGAGAGGC
CGCGTTCAGTCGGGTCCCGGCAGCGGCTGCAGCGCTCTCGTCTTCTGCGGCTCTCGGTGCCCTCTCCTTTTCGTT

TCCGGAAACATGGCCTCCGGTGTGGCTGTCTCTGATGGTGTCATCAAGGTGTTCAACGACATGAAGGTGCGTAAG
TCTTCAACGCCAGAGGAGGTGAAGAAGCGCAAGAAGGCGGTGCTCTTCTGCCTGAGTGAGGACAAGAAGAACATC
ATCCTGGAGGAGGGCAAGGAGATCCTGGTGGGCGATGTGGGCCAGACTGTCGACGACCCCTACGCCACCTTGTCA
AGATGCTGCCAGATAAGGACTGCCGCTATGCCCTCTATGATGCAACCTATGAGACCAAGGAGAGCAAGAAGGAGG
ATCTGGTGTTTATCTTCTGGGCCCCCGAGTCTGCGCCCCTTAAGAGCAAAATGATTTATGCCAGCTCCAAGGACG
CCATCAAGAAGAAGCTGACAGGGATCAAGCATGAATTGCAAGCAAACTGCTACGAGGAGGTCAAGGACCGCTGCA
CCCTGGCAGAGAAGCTGGGGGGCAGTGCCGTCATCTCCCTGGAGGGCAAGCCTTTGTGAGCCCCTTCTGGCCCCC
TGCCTGGAGCATCTGGCAGCCCCACACCTGCCCTTGGGGGTTGCAGGCTGCCCCCTTCCTGCCAGACCGGAGGGG
CTGGGGGGATCCCAGCAGGGGGAGGGCAATCCCTTCACCCCAGTTGCCAAACAGACCCCCCACCCCCTGGATTTT
CCTTCTCCCTCCATCCCTTGACGGTTCTGGCCTTCCCAAACTGCTTTTGATCTTTTGATTCCTCTTGGGCTGAAG
CAGACCAAGTTCCCCCCAGGCACCCCAGTTGTGGGGGAGCCTGTATTTTTTTTAACAACATCCCCATTCCCCACC
TGGTCCTCCCCCTTCCCATGCTGCCAACTTCTAACCGCAATAGTGACTCTGTGCTTGTCTGTTTAGTTCTGTGTA
TAAATGGAATGTTGTGGAGATGACCCCTCCCTGTGCCGGCTGGTTCCTCTCCCTTTTCCCCTGGTCACGGCTACT
CATGGAAGCAGGACCAGTAAGGGACCTTCGATTAAAAAAAAAAAAGACAATAATAAAAA

Protéine issue de l'ARNm infidèle

MASGVAVSDGVIKVFNDMKVRKSSTPEEVKKRKKAVLFCLSEDKKNIILEEGKEILVGDVGQTVDDPYATLSRCC
QIRTAAMPSMMQPMRPRRARRRIWCLSSGPPSLRPLRAK

ATG Codons start et stop.

TTT Répétition de bases contenant la position de délétion.

MAS Partie protéique commune entre la protéine de référence et la protéine aberrante.

LSR TIP.

Figure 61 : Étapes de la prédiction d'un TIP exemplifié par la délétion 447 de CFL1.

4.3.2 Prédiction des TIPs (Transcription Infidelity Peptides)

Les validations biologiques ayant débuté après l'obtention des données 2007, les délétions et peptides présentés dans cette partie correspondent aux résultats bioinformatiques de 2007.

Nous avons, en 2007, défini un ensemble de 2.761 positions ARNm pour lesquelles la proportion d'ESTs cancéreuses portant une délétion est significativement plus grande que la proportion d'ESTs normales portant la même délétion (délétions C>N obtenues par application du test de proportions et après application du filtre -10/+10).

2.194 de ces positions sont situées dans la partie codante et permettent donc de définir un ensemble de 2.194 TIPs.

4.3.3 Formulation de l'hypothèse biologique

Les prédictions bioinformatiques suggèrent donc la production de protéines porteuses de fragments aberrants, *i.e.* non codés par le génome.

Hypothèse de travail : la traduction d'un ARNm présentant une délétion au sein de la partie codante aboutit probablement à une protéine dont la structure secondaire et/ou tertiaire est

altérée. Ces protéines sont normalement dégradées par le protéasome. Les peptides générés par cette dégradation peuvent présenter un caractère immunogénique. Cette hypothèse peut être validée biologiquement en recherchant directement les anticorps dirigés contre les TIPs prédits par la bioinformatique.

L'hypothèse bioinformatique d'infidélité de transcription est validée biologiquement au niveau de l'ARN et au niveau de la protéine. Par ailleurs, l'idée originale développée par Genclis est la recherche non pas des protéines aberrantes directement, mais des anticorps générés en réponse à la présence de protéines aberrantes. Se pose alors la question de savoir si cette découverte a le potentiel d'aider au diagnostic des cancers.

5 Utilisation des TIPs pour le diagnostic des cancers

5.1 Détection d'anticorps dirigés contre les TIPs

5.1.1 TIPs sélectionnés pour la validation biologique

60 TIPs ont été sélectionnés sur plusieurs critères :

- ✓ La délétion générant le TIP ne correspond pas à une position de SNP référencée dans dbSNP.
- ✓ Il n'existe pas de SNP décalant le cadre de lecture en amont de la délétion et générant le même peptide aberrant d'après les données dbSNP du NCBI.

Remarque : la banque dbSNP est mise à jour et consultée régulièrement. Néanmoins, les données présentées dans ce manuscrit sont susceptibles d'évoluer en fonction des nouvelles données déposées.

Il est important de rappeler ici que les SNPs déposés dans dbSNP sont validés biologiquement ou non. En décembre 2008, 20 millions de SNPs humains sont référencés, dont 50% seulement sont validés biologiquement. Un SNP non validé biologiquement peut venir d'études à grande échelle d'ESTs. Dans ce cas et au vu des résultats précédemment présentés, on ne peut affirmer s'il s'agit de variations de l'ADN (SNP) ou de l'ARN (IT).

- ✓ Longueur du TIP ≥ 15 AA.
- ✓ Pas d'identité de plus de 7 AA consécutifs entre le TIP et toute protéine humaine après analyse des alignements des TIPs contre l'ensemble des protéines humaines.

Les délétions générant les TIPs sélectionnés sont représentatives de l'ensemble des délétions en termes de pourcentages de déviation dans les ESTs normales et cancéreuses.

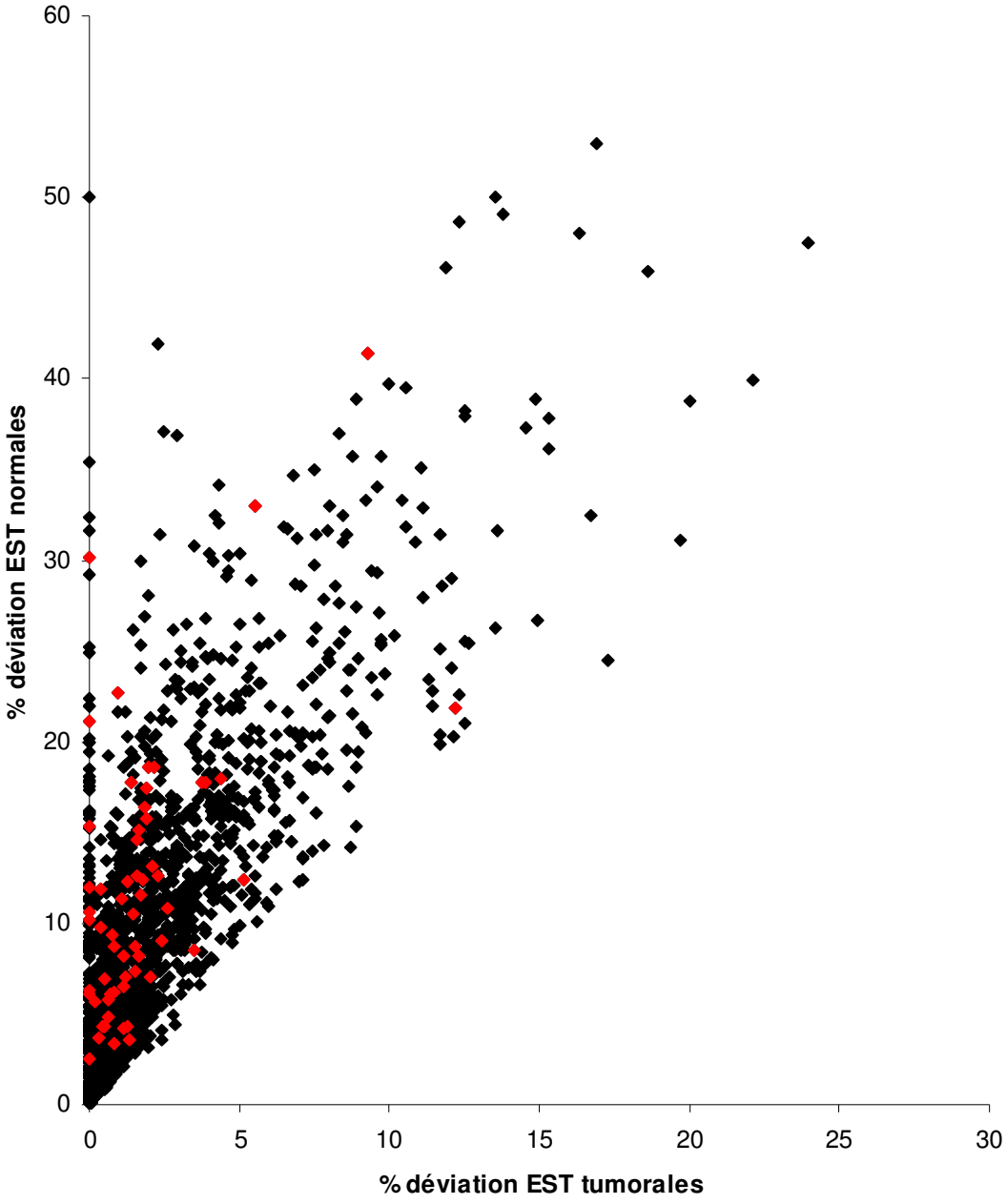


Figure 62 : Choix de 60 délétions parmi l'ensemble des gaps C>N.

Chaque point représente une délétion C>N, caractérisée par le pourcentage de déviation des ESTs cancéreuses en fonction du pourcentage de déviation des ESTs normales. Les losanges rouges représentant les 60 délétions sélectionnées.

Pour des raisons techniques, les séquences des 60 TIPs ainsi obtenues sont conservées en totalité si elles sont inférieures à 30 AA et en général coupées à 30 AA dans le cas contraire.

5.1.2 Principe du test

Le principe du test biologique est relativement classique (principe de l'ELISA) puisqu'il s'agit de détecter dans le sérum humain la présence d'anticorps dirigés contre des TIPs servant d'appâts. Ces anticorps sont només TIAB (Transcription Infidelity Antibodies).

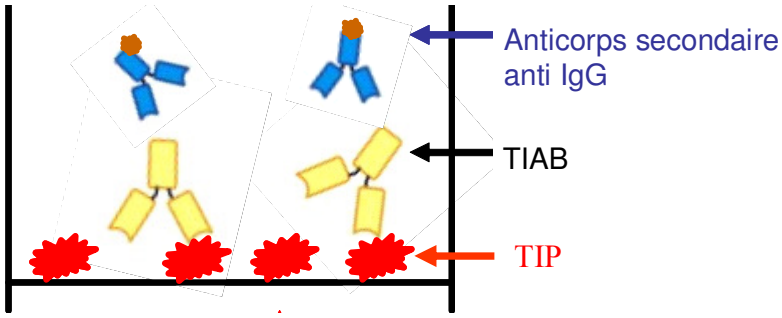


Figure 63 : Principe du test ELISA.

5.1.3 Mise en place de contrôles négatifs

5.1.3.1 Choix des contrôles négatifs

Un peptide dit canonique ou CP est défini comme étant le peptide codé par la même séquence ARNm qu'un TIP, mais dans le cadre de lecture de référence. Il s'agit donc d'un peptide normal, devant être reconnu comme un élément du soi. Il ne doit par conséquent pas entraîner la production d'anticorps.

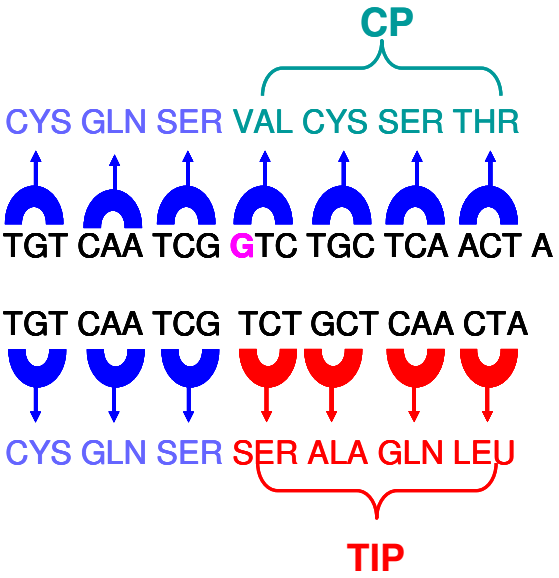


Figure 64 : Représentation schématique d'un CP.

Trois peptides canoniques correspondant aux TIPs 7, 24 et 28 ont été sélectionnés.

5.1.3.2 Test des contrôles négatifs

Les 60 TIPs ainsi que les trois peptides canoniques sont testés sur 20 individus contrôles provenant de l'Établissement Français du Sang (EFS). Le test de Wilcoxon apparié montre que chacun des 60 TIPs testés présente des valeurs significativement plus hautes que chacun des trois peptides canoniques. Par ailleurs, pour 41 TIPs, chacun des 20 témoins présente des valeurs plus élevées pour le TIP que pour chacun des trois peptides canoniques. Ainsi, en prenant le point de vue le plus stringent, les prédictions bioinformatiques sont correctes avec une précision de 68%.

Le choix des TIPs testés est donc réalisé de manière bioinformatique. Les premiers dosages montrent que les peptides canoniques prédits comme contrôles négatifs donnent des résultats négatifs. Ils montrent également que les prédictions des 60 TIPs sont correctes. Ces TIPs sont alors exploités pour discriminer les patients cancéreux des individus non porteurs de cancer. Les dosages sont bien sûr réalisés par les équipes de biologistes et l'analyse des résultats revient à l'équipe statistique.

5.2 Discrimination sérum cancéreux / sérum non cancéreux

Le screening réalisé sur 20 patients atteints de cancer du poumon et 20 sujets témoins a permis de sélectionner en sous-ensemble de 6 TIPs potentiellement discriminants. Ces 6 TIPs sont testés sur deux populations de plus grande taille, à savoir 133 témoins de l'EFS et 131 patients atteints de cancer du poumon ayant été pris en charge par le CHU de Grenoble.

Le cancer du poumon a été choisi pour cette première étude. En effet, la cancer du poumon est le plus mortel en France car il est le plus fréquent et, en même temps, il présente un taux de survie à 5 ans de seulement 12% [148]. Par ailleurs, il n'existe pas à l'heure actuelle de test de diagnostic pertinent du cancer du poumon. Plusieurs marqueurs tumoraux sont décrits, comme CEA (carcino embryonic antigen) ou CYFRA 21-1, qui présentent une spécificité de 91% pour une sensibilité de 42% et 54% respectivement [149]. L'examen majeur en cas de suspicion de cancer reste la radiographie du thorax. Le traitement par chirurgie donne les meilleurs résultats sur des cancers détectés au stade précoce, ce qui représente seulement 30% des diagnostics en France. Il est donc de toute première importance de développer des outils permettant le diagnostic précoce du cancer du poumon.

L'étude présentée se concentre sur les cancers du poumon non à petites cellules (NSCLC pour non small-cell lung carcinomas), qui représentent 85% des cancers bronchiques. Les types histologiques de cancers du poumon non à petites cellules de cette population sont les adénocarcinomes (47%), les cancers épidermoïdes (29%) et les autres types histologiques regroupés dans une troisième catégorie (24%).

5.2.1 Données cliniques

Les témoins et patients cancéreux sont appariés en âge et en sexe. L'appariement hommes / femmes est évident, l'appariement en âge est basé sur une méthode de stratification. En d'autres termes, les quartiles sont équilibrés dans les deux populations que l'on étudie.

	Nombre	Age	Hommes	Femmes
Témoins	133	58 ± 6	103	30
NSCLC	131	61 ± 11	102	29
T+N0M0	79	61 ± 11	64	15
T+N+M0	40	61 ± 12	27	13
T+N+M+	12	62 ± 11	11	1
Adénocarcinomes	62	61 ± 9	42	20
Epidermoïdes	38	65 ± 11	34	4
Autres	31	58 ± 13	26	5

Table 23 : Tableau clinique des témoins et patients atteints de cancer du poumon.

Cette population d'étude n'est pas représentative de la pathologie du cancer du poumon dans la mesure où elle comprend 60% de diagnostics précoces. Ce choix est justifié par notre objectif premier, qui est la détection précoce des cancers.

5.2.2 Données brutes

Les mesures d'anticorps dirigés contre les 6 TIPs sélectionnés sont réalisées par un test ELISA comme décrit précédemment. Les mesures sont réalisées de manière automatisée. Les valeurs des anticorps dirigés contre les 6 TIPs d'intérêt sont mesurées simultanément. N'ayant pas de standard interne, les valeurs sont sommées puis normalisées. En d'autres termes, pour chaque individu, la valeur mesurée contre chaque TIP est ramenée à sa participation relative à la somme des 6 valeurs mesurées.

NCBI nom du gène	IgG anti-TIP (mean ± SEM)		P-valeurs	
	Contrôles	*NSCLC*	test F	test T
TIP#1	0,120 ± 0,002	0,152 ± 0,002	5×10^{-4}	5×10^{-25}
TIP#2	0,084 ± 0,001	0,115 ± 0,003	2×10^{-14}	2×10^{-20}
TIP#3	0,062 ± 0,001	0,076 ± 0,001	0,11	5×10^{-12}
TIP#4	0,141 ± 0,003	0,163 ± 0,002	0,11	6×10^{-9}
TIP#5	0,299 ± 0,003	0,236 ± 0,003	0,29	9×10^{-37}
TIP#6	0,294 ± 0,003	0,258 ± 0,003	0,12	4×10^{-15}

Table 24 : Résultats des dosages de 6 TIPs sur l'ensemble des témoins et des patients atteints de cancer du poumon décrits table 23.

Ainsi, les valeurs des anticorps dirigés contre les peptides aberrants TIP#1, #2, #3 et #4 sont augmentées chez les patients atteints de cancer du poumon par rapport aux sujets contrôles. Les valeurs des anticorps dirigés contre les peptides aberrants TIP#5 et #6 sont au contraire abaissées.

Il est important de noter que ces variations ne sont pas dues à la normalisation des données.

5.2.3 Analyse statistique des résultats

Plusieurs règles de classement statistique ont été évaluées afin de discriminer les sujets sains des sujets atteints de cancer.

Les cinq méthodes utilisées sont décrites brièvement de la manière suivante :

- ✓ Support Vector Machine (SVM)

Les individus sont représentés géométriquement dans l'espace à *n* dimensions, où *n* est le nombre de variables explicatives. Dans cette étude, *n* est le nombre de TIPs utilisés pour établir la règle de classement. Les SVM sont une règle géométrique qui permet de déterminer l'hyperplan de séparation qui minimise le nombre d'individus mal classés. L'hyperplan est une droite pour deux variables, un plan pour trois variables, …

Les individus sont affectés à l'une ou l'autre des classes (ici, cancer ou absence de cancer) en fonction de leur position par rapport à l'hyperplan.

Exemple de SVM avec 2 variables :

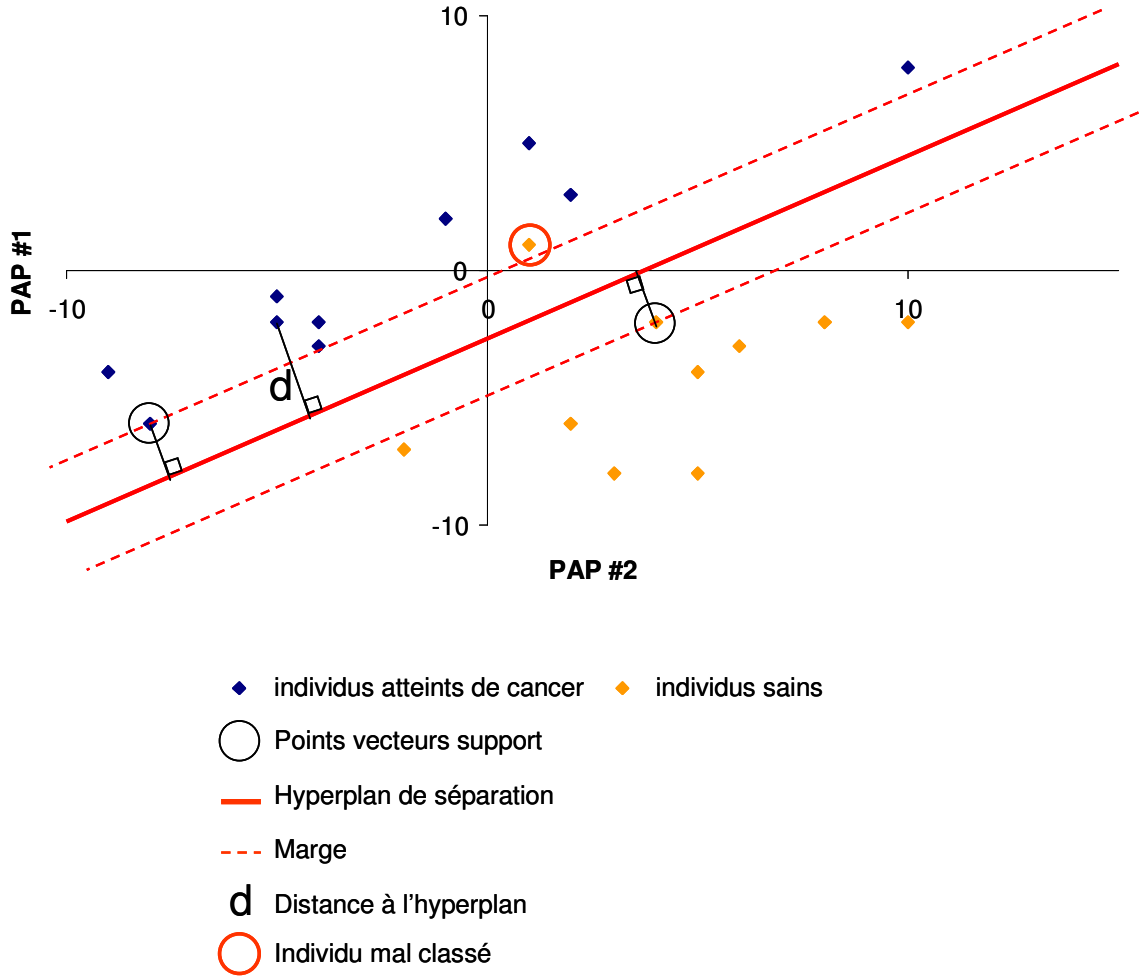


Figure 65 : Représentation schématique d'une SVM.

✓ Modèle logistique

Cette méthode probabiliste permet de calculer, pour chaque individu, la probabilité d'appartenir à l'une ou l'autre des deux classes. Le logarithme de chaque probabilité dépend linéairement d'une série de paramètres qui correspondent aux variables explicatives. Ces paramètres sont estimés et leur nullité est testée ; les paramètres significativement nuls sont éliminés du modèle. Chaque individu est affecté à la classe de probabilité maximale.

✓ Règle de classement linéaire

Les individus sont représentés géométriquement dans l'espace à *n* dimensions, où *n* est le nombre de variables explicatives. Cette méthode consiste à affecter un individu à la classe dont il est le plus proche du centre.

✓ Règle de classement quadratique

Contrairement à la règle de classement linéaire, la dispersion de chaque classe est ici prise en compte dans le calcul de la distance.

✓ Arbres de décision

La méthode détermine à chaque pas la variable et son découpage qui permettent de créer des feuilles les plus éloignées possible au sens d'une distance à choisir. La méthode s'arrête quand aucune autre séparation n'améliore le classement. D'autres critères comme le nombre de pas ou le nombre minimal d'individus dans les feuilles peuvent être fixés comme critères d'arrêt.

Les différentes méthodes sont testées et leur robustesse est évaluée par *leave-one-out*. Le *leave-one-out* consiste à établir la règle de classement sur N-1 individus (N étant le nombre total d'individus à classer, soit 264). L'individu restant est classé à postériori. Cette démarche est appliquée successivement à l'ensemble des individus.

Ces cinq méthodes ont été appliquées au jeu de données comprenant 133 témoins et 131 patients atteints de cancer du poumon. Les performances des différentes règles de classement sont données ci-dessous :

	Resubstitution		**Leave-one-out**	
	Specificité	**Sensitibilité**	**Specificité**	**Sensitibilité**
Support Vector Machine	95%	84%	95%	83%
Modèle logistique	95%	84%	94%	84%
Analyse discriminante linéaire	93%	82%	93%	81%
Analyse discriminante quadratique	92%	81%	92%	81%
Arbre de décision	90%	85%	90%	83%

Table 25 : Résultat de l'étude discriminante entre témoins et patients atteints de cancer du poumon par différentes méthodes de classement.

Il est important de noter que les cinq méthodes sont très cohérentes entre elles, dans la mesure où toutes donnent une spécificité supérieure à 90% et une sensibilité supérieure à 80%.

La représentation des résultats obtenue par SVM est réalisée par l'intermédiaire de la distance algébrique à l'hyperplan. Le seuil de positivité est fixé par défaut à zéro.

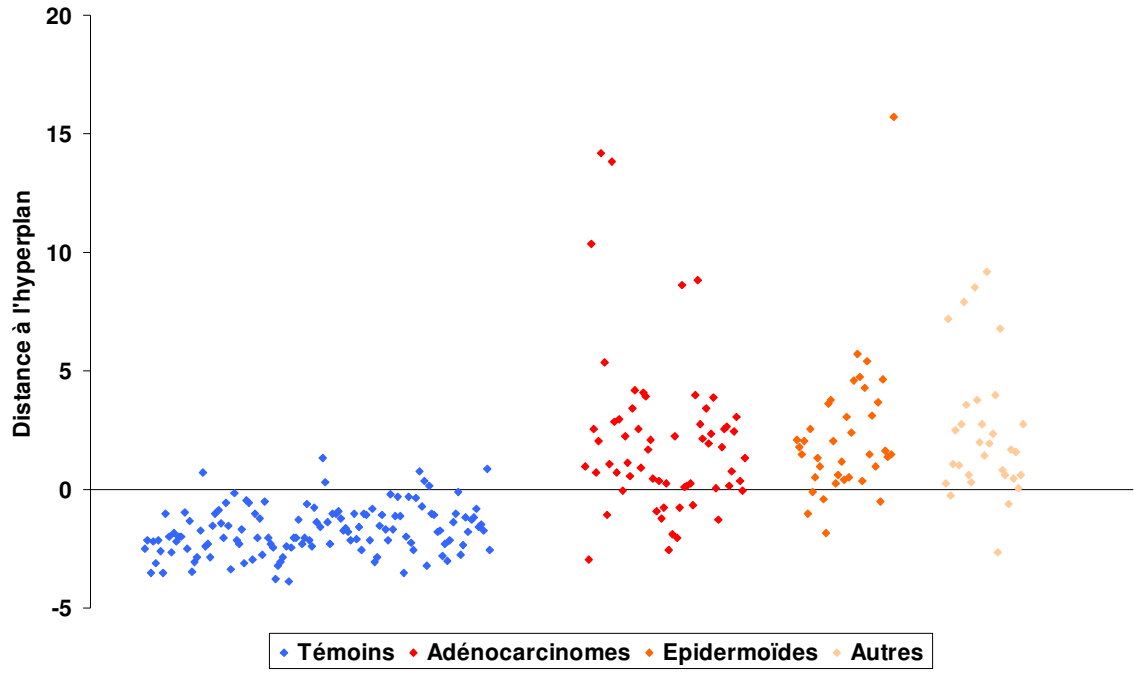


Figure 66 : Représentation de la distance algébrique à l'hyperplan obtenue par une SVM basée sur 5 TIPs.

Il est donc possible de discriminer patients cancéreux et témoins à l'aide des anticorps dirigés contre des peptides issus d'évènements d'infidélité de transcription. Afin de mieux comprendre les mécanismes mis en jeu lors de l'évolution de la maladie, les recherches se sont également orientées vers le modèle murin.

5.3 Perspectives d'étude chez Mus musculus

5.3.1 Analyse des ESTs de *Mus musculus*

L'analyse des ESTs souris n'a pu être réalisée que tardivement. En effet, l'analyse repose sur la comparaison des ESTs issues de tissus cancéreux aux ESTs issues de tissus non cancéreux. Or, la source cancéreuse ou non des librairies d'ESTs murines n'a été rendue disponible sur le NCBI qu'en novembre 2008. L'analyse a donc été entreprise à cette date.

Les données ESTs de *Mus musculus* sont biaisées en termes d'effectifs, dans la mesure où l'on dispose de 4 millions d'ESTs d'origine non cancéreuse pour 460.000 ESTs d'origine cancéreuse.

	Homo sapiens	***Mus musculus***
ESTs cancer	2.608.711	460.300
ESTs normal	2.985.560	3.990.021

Table 26 : Nombre d'ESTs disponibles pour l'homme et la souris.

L'étude est réalisée de la même manière que l'exploitation des ESTs humaines.

Résultats bruts :

Substitutions		C>N	LBE	N>C	LBE	CndN	(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	1620	589	4534	246	2201	0,4	0,2
	Fisher	2591	2427	6372	1219	23359	0,4	0,0
	somme	4211	3017	10906	1465	25560	0,4	0,1

Délétions		C>N	LBE	N>C	LBE	CndN	(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	663	49,1	250	86,9	447	2,7	3,8
	Fisher	1281	169	331	259	2256	3,9	15,4
	somme	1944	218	581	346	2703	3,3	7,3

Insertions I		C>N	LBE	N>C	LBE	CndN	(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	122	6,88	28	16,9	88	4,4	10,4
	Fisher	297	19,1	40	49,3	292	7,4	ND
	somme	419	26	68	66,2	380	6,2	224,2

Insertions N		C>N	LBE	N>C	LBE	CndN	(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	143	13,2	32	25,6	213	4,5	20,2
	Fisher	302	70,3	67	110	1203	4,5	ND
	somme	445	83,5	99	136	1416	4,5	ND

Insertions T		C>N	LBE	N>C	LBE	CndN	(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	331	21,5	67	50,3	320	4,9	18,5
	Fisher	804	104	106	206	1829	7,6	ND
	somme	1135	126	173	256	2149	6,6	ND

Table 27 : Exploitation des données non filtrées. Nombre de positions présentant un test (Fisher ou proportions) significatif ou non pour chaque type d'événement.

Le nombre de tests réalisés et significatifs est plus faible que chez l'homme.

Il est important de soulever une différence notable entre les résultats humains et murins. Cette différence concerne les substitutions ; l'étude à grande échelle des ESTs humaines a montré que les substitutions de base surviennent significativement plus fréquemment dans les ESTs cancéreuses. L'étude des ESTs murines montre au contraire un ratio (C>N) / (N>C) inférieur à 1, ce qui signifie que le nombre de positions où la proportion de substitutions est

significativement plus importante dans les ESTs normales que dans les ESTs cancéreuses (positions N>C) est plus grand que le nombre de positions C>N.

Notons également que les évènements d'insertion existent chez la souris mais que le nombre de positions présentant un test significatif est faible. Néanmoins, les ratios (C>N) / (N>C) des insertions sont très largement supérieurs à 1.

Enfin, les délétions C>N existent chez la souris avec des effectifs nettement plus grands que les délétions N>C.

Résultats après application du filtre -10/+10 :

Substitutions		C>N	LBE	N>C	LBE	CndN	(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	701	128	989	89	486	0,7	0,6
	Fisher	1070	593	1702	327	5472	0,6	0,3
	somme	1771	721	2691	416	5958	0,7	0,5

Délétions		C>N	LBE	N>C	LBE	CndN	(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	251	15,7	97	30	109	2,6	3,5
	Fisher	585	64,3	137	101	793	4,3	14,6
	somme	836	80	234	131	902	3,6	7,4

Insertions I		C>N	LBE	N>C	LBE	CndN	(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	11	0,35	1	1,15	3	11,0	ND
	Fisher	54	2,89	2	9,27	55	27,0	ND
	somme	65	3,24	3	10,4	58	21,7	ND

Insertions N		C>N	LBE	N>C	LBE	CndN	(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	8	1,13	7	0,97	6	1,1	1,1
	Fisher	64	14,2	12	21,7	236	5,3	ND
	somme	72	15,3	19	22,7	242	3,8	ND

Insertions T		C>N	LBE	N>C	LBE	CndN	(C>N) / (N>C)	[(C>N) - LBE] / [(N>C) - LBE)]
	Chi²	25	1,62	7	2,98	14	3,6	5,8
	Fisher	142	19,3	11	38,3	353	12,9	ND
	somme	167	20,9	18	41,3	367	9,3	ND

Table 28 : Exploitation des données filtrées. Nombre de positions présentant un test (Fisher ou proportions) significatif ou non pour chaque type d'événement.

Comme chez l'homme, l'impact du filtre est notable et différent suivant le type d'événement.

La proportion de substitutions C>N perdues après application du filtre est moins grande que pour les substitutions N>C. Le nombre de tests significatifs est donc réduit, mais le ratio (C>N) / (N>C) tend à s'approcher de 1.

Le filtre a également un effet important sur les délétions ; néanmoins, le ratio (C>N) / (N>C) reste élevé.

Enfin, l'impact du filtre sur les insertions est très fort et le nombre d'événements résistant à l'application du filtre est faible.

5.3.2 Identification d'évènements homologues

L'infidélité de transcription affectant l'homme et la souris, nous avons cherché à savoir si les positions des évènements étaient communes entre les deux espèces.

Au vu des résultats de l'étude des ESTs de *Mus musculus*, les positions issues de tests de proportions et de tests de Fisher ont été regroupées. Le travail se concentre par ailleurs sur les données filtrées.

La base Homologene [150] du NCBI a été exploitée de manière à identifier les gènes orthologues. Les transcrits des deux espèces sont ensuite alignés de manière à déterminer la correspondance entre les positions sur le transcrit humain et les positions sur le transcrit murin. Par exemple, plus de la moitié des transcrits portant une délétion C>N chez la souris présentent un homologue humain portant lui aussi au moins une délétion.

			Homo sapiens	Mus musculus	communs
Substitutions	C>N	nombre de transcrits	988	739	223
		nombre de positions	3738	1771	24
	N>C	nombre de transcrits	849	1307	138
		nombre de positions	1813	2691	7
Délétions	C>N	nombre de transcrits	1247	264	135
		nombre de positions	5645	836	173
	N>C	nombre de transcrits	273	176	33
		nombre de positions	484	234	4
Insertions	C>N	nombre de transcrits	409	69	19
		nombre de positions	2566	167	17
	N>C	nombre de transcrits	300	16	2
		nombre de positions	501	18	0

Table 29 : Nombre de transcrits et positions C>N ou N>C homologues entre homme et souris.

La figure ci-dessous montre la particularité des délétions C>N, pour lesquelles plus de 20% des positions C>N sont retrouvées sur le même transcrit et à la même position chez l'homme et la souris.

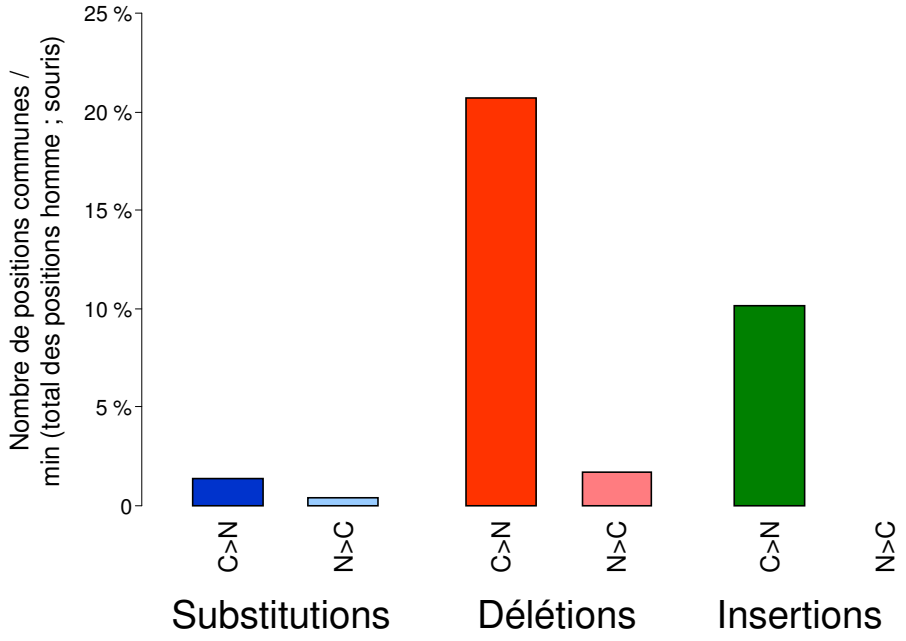


Figure 67 : Représentation des positions communes C>N et N>C de substitutions, délétions et insertions.

La pertinence statistique de cette observation a ensuite été étudiée. Pour chaque couple de transcrits homologues, on dispose de l'alignement des transcrits qui nous permet de déterminer la longueur de la partie commune. Par ailleurs, on dispose du nombre de positions statistiquement significatives et contenues dans la partie alignée pour l'homme et la souris, ainsi que du nombre *N* de positions communes aux deux espèces. Il est alors possible de calculer, pour chaque couple de transcrits, la probabilité d'observer, par hasard, *N* ou plus de *N* positions communes.

Voici les résultats obtenus pour les transcrits orthologues présentant plus de 3 positions de gaps C>N filtre communes :

Identifiants des transcrits Hs et Mm	Taille de la partie homologue des transcrits Hs et Mm	Nombre de positions Hs	Nombre de positions Mm	Nombre de positions communes (k)	P(PC≥k)
NM_001614 / NM_009609	1621	58	19	14	1,1E-17
NM_006597 / NM_031165	1839	28	17	13	2,0E-22
NM_001404 / NM_026007	1398	47	11	10	7,3E-15
NM_001428 / NM_023119	1445	34	9	7	7,3E-11
NM_001416 / NM_144958	1579	14	6	6	1,4E-13
NM_006098 / NM_008143	978	28	9	6	2,5E-08
NM_001005 / NM_012052	751	26	9	6	7,4E-08
NM_000291 / NM_008828	1575	14	6	5	1,5E-10
NM_001013 / NM_029767	587	13	6	5	1,3E-08
NM_006082 / NM_011654	1449	45	7	5	4,6E-07

NM_000968 / NM_024212	1190	32	9	5	1,2E-06
NM_005324 / NM_008211	1608	10	6	4	1,1E-08
NM_001686 / NM_016774	1535	13	10	4	6,3E-07
NM_000967 / XM_001477264	1150	29	10	4	6,2E-05
NM_006136 / NM_007604	1715	6	3	3	2,4E-08
NM_001970 / NM_181582	1111	5	5	3	4,4E-07
NM_002635 / NM_133668	1297	12	3	3	6,1E-07
NM_002568 / NM_008774	2675	11	6	3	1,0E-06
NM_004261 / NM_053102	1154	7	6	3	2,7E-06
NM_014713 / NM_008640	1187	9	6	3	6,0E-06
NM_000942 / NM_011149	810	15	4	3	2,0E-05
NM_020300 / NM_019946	661	7	7	3	2,5E-05
NM_001007 / NM_009094	770	24	3	3	2,7E-05

Table 30 : Pertinence statistique des homologies homme / souris.

Les résultats obtenus sur les homologues homme / souris sont donc, pour les gaps C>N, largement au dessus des probabilités d'observer ces mêmes résultats par hasard.

5.3.3 Preuve de concept *in vivo*

Les expériences réalisées chez la souris ont consisté à injecter des cellules de cancer du poumon, Lewis lung carcinomas (LLC1), à des souris C57Bl/6 qui développent ainsi une tumeur pulmonaire, puis à mesurer les anticorps dirigés contre 3 TIPs hautement conservés entre *Homo sapiens* et *Mus musculus*.

Un peptide canonique parfaitement conservé entre homme et souris est utilisé comme contrôle négatif.

Les mesures sont réalisées le jour de l'injection puis, 7, 14 et 21 jours après l'injection de cellules cancéreuses sur 10, 16 et 22 souris.

Les premiers résultats montrent que les valeurs des anticorps dirigés contre les peptides aberrants sont augmentées après trois voire deux semaines suivant l'injection des cellules LLC1. Aucun signal n'est détecté pour le contrôle négatif.

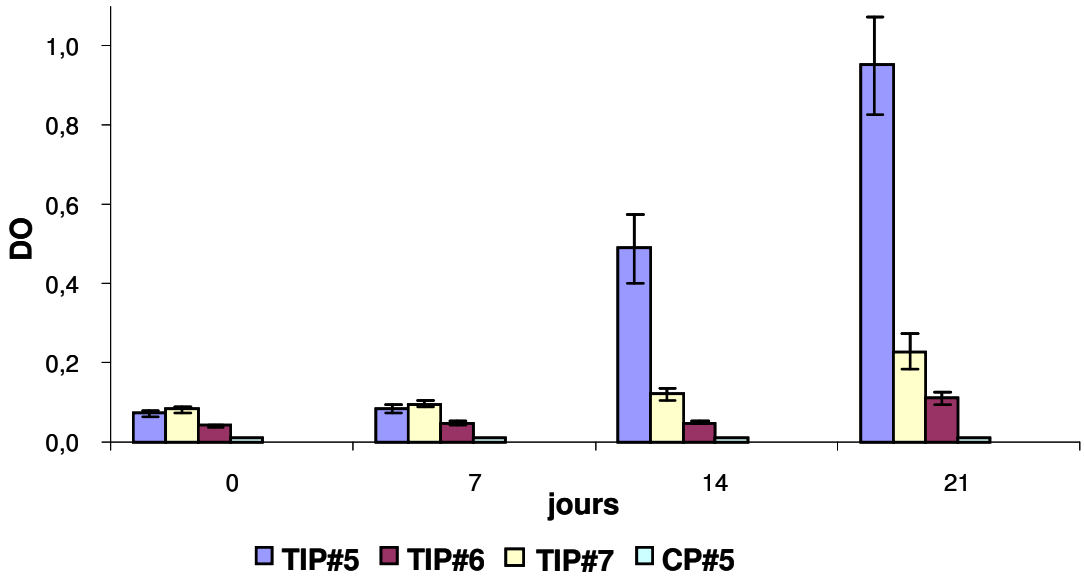


Figure 68 : Résultats des dosages de 3 TIPs et d'un CP.

Les 3 TIPs sont représentés en violet, jaune et fuchsia. Le contrôle est représenté en bleu clair.

Les premiers résultats *in vivo* montrent donc que le phénomène existe chez la souris comme chez l'homme, ce qui nous offre de grandes perspectives d'étude des mécanismes liés à l'infidélité de transcription.

6 Conclusions et perspectives

Le travail décrit dans ce manuscrit consiste en une analyse des séquences ESTs représentant les ARN des tissus cancéreux et non cancéreux. Les principales conclusions de cette analyse bioinformatique sont les suivantes :

- ✓ Il existe une augmentation de l'infidélité de transcription dans les cellules tumorales par rapport aux cellules normales.
- ✓ Trois types d'évènements sont observés, à savoir les substitutions, les délétions et les insertions de bases.
- ✓ Les évènements d'infidélité de transcription observés peuvent être considérés comme des phénomènes normaux dont l'intensité augmente dans la pathologie cancéreuse.
- ✓ L'infidélité de transcription est conditionnée par le contexte d'ADN sur une distance correspondant à la partie d'ADN ouverte sous l'action de l'ARNpol II.
- ✓ La survenue des évènements d'infidélité de transcription est plus fréquente que la survenue de mutations somatiques.

Cette thèse n'est pas le premier travail évoquant l'infidélité de transcription, mais l'originalité de notre contribution réside d'une part dans la découverte de l'accroissement du phénomène dans le cancer et d'autre part dans la réalisation de différentes validations biologiques basées sur des prédictions bioinformatiques. Les principales inférences biologiques validées sont les suivantes :

- ✓ Le répertoire protéique humain est vraisemblablement beaucoup plus diversifié qu'attendu sur la base des 40.000 gènes que comporte notre génome. Néanmoins, n'ayant pu réaliser une analyse statistique que sur 0,2% de l'ensemble des positions du transcriptome humain, il n'est pas possible d'évaluer le nombre de variants protéiques issus de la traduction d'ARN infidèles.

  Nous avons démontré l'existence d'un variant protéique issu de la substitution du codon stop en triplet codant, à savoir l'ApoAII-PSP, chez tous les sujets normaux testés. Cette découverte nous a permis de mettre en évidence une lacune importante dans la méthodologie d'interprétation des données de spectrométrie de masse. En effet, il a été nécessaire de créer une nouvelle banque de données contenant la séquence peptidique de l'ApoAII-PSP pour obtenir une identification du variant protéique. Sans cette nouvelle banque, les spectres obtenus auraient été éliminés

puisque leur interprétation n'était pas compatible avec une lecture canonique du gène de l'ApoAII. Un travail d'annotation bioinformatique de l'ensemble des séquences ARNm permettra de créer un protéome alternatif dont il faudra ensuite valider l'existence.

La présence du variant ApoAII-PSP chez tous les sujets normaux est établie. Ceci indique que les protéines variantes ne sont pas systématiquement dégradées. Toutefois, on ne peut exclure la possibilité que ces variants soient la conséquence d'une erreur survenant au moment de la traduction. La seule conclusion que nous pouvons tirer est donc qu'une lecture canonique du génome humain est insuffisante pour expliquer la complexité du répertoire protéique humain normal et pathologique.

La principale conséquence de cette découverte réside dans la prise de conscience de l'extrême diversité du protéome. Ainsi, les études de protéomique du cancer manquent de reproductibilité, non pour des raisons techniques, mais plus vraisemblablement parce que la taille de l'échantillon analysé n'est pas adéquate au vu de l'hétérogénéité protéique.

- ✓ Il existe des ARNm portant une délétion dans les cellules cancéreuses et ce, en l'absence de mutation somatique. Nous avons validé biologiquement cette prédiction bioinformatique par le séquençage d'un transcrit infidèle issu du gène CFL1. L'absence d'une base au niveau de la partie codante d'un ARN a pour conséquence le décalage du cadre de lecture. Une partie des ARN variants issus de la délétion d'une base au cours de la transcription est probablement dégradée grâce à l'intervention des différents mécanismes de surveillance des ARN. Les produits de dégradation des protéines issues de la traduction des ARN échappant aux mécanismes de surveillance pourraient servir de stimuli immunogéniques.

  Ayant constaté que l'approche protéomique des biomarqueurs du cancer représentait une tâche titanesque, voire insurmontable, compte tenu des technologies actuelles, nous nous sommes tournés vers une démonstration indirecte de l'existence des biomarqueurs. La présence d'anticorps dirigés spécifiquement contre des peptides aberrants issus d'ARNm infidèles a été démontrée chez l'homme et chez la souris. La spécificité de ces anticorps est établie par l'absence de réactivité des séras humains et murins contre les peptides canoniques. Rappelons que les immunoglobulines dirigées contre les peptides aberrants sont présentes chez les individus témoins ; ceci conforte l'idée que l'infidélité de transcription est un phénomène normal de faible intensité qui induit une grande diversité du répertoire protéique.

Ainsi, certaines protéines issues d'ARN infidèles sont sécrétées, comme l'ApoAII ; d'autres sont dégradées et possèdent des propriétés immunogéniques. L'infidélité de transcription pourrait donc apporter une explication à l'origine des immunoglobulines naturelles. En effet, l'existence des immunoglogulines naturelles est connue et leur rôle physiologique est de servir de première ligne de défense contre les infections. Toutefois, la nature des ligands de ces immunoglobulines était jusqu'à présent inconnue. L'étude de l'infidélité de transcription a permis d'identifier des peptides aberrants ; ces peptides représentent vraisemblablement les épitopes reconnus par les immunoglobulines naturelles. Cette découverte ouvre la voie d'un diagnostic précoce des cancers.

Dans une perspective très différente, nous avons montré que le contexte d'ADN exerçait un rôle important sur la survenue d'évènements d'infidélité de transcription. Ainsi, les substitutions sont conditionnées par les sept bases correspondant à la partie d'ADN ouverte sous l'action de l'ARNpol II au cours de la transcription. Il est possible que ces structures transitoires permettent le glissement de l'enzyme, expliquant l'incorporation erronée de la base précédant l'événement. Par ailleurs, l'étude de l'ensemble du transcriptome humain a montré que la base affectée par un événement de substitution était préférentiellement localisée en phase 3 du codon. Cette observation est statistiquement significative. Or, le concept de codon devient une réalité biologique seulement au moment de la traduction. D'autres analyses doivent être développées afin d'expliquer ce phénomène.

Des travaux développés parallèlement ont montré que les évènements d'infidélité de transcription ne sont pas indépendants les uns des autres. En effet, un événement d'infidélité de transcription sur une EST peut être dépendant de la survenue d'évènements préalables. L'influence potentielle des structures secondaires de l'ARNpm reste à définir. Ainsi, le travail présenté dans ce manuscrit ouvre différentes perspectives puisqu'il fournit les prémices d'un code porté par l'ADN et allant bien au delà d'un appariement de bases tel que défini par Watson et Crick.

Enfin, nous avons montré l'importance du contexte d'ADN sur les mutations somatiques autres que les transitions C : G → A : T. Il est par conséquent concevable que les mutations somatiques naissent de certains évènements d'infidélité de transcription. L'hypothèse mécanistique proposée peut être représentée de la manière suivante :

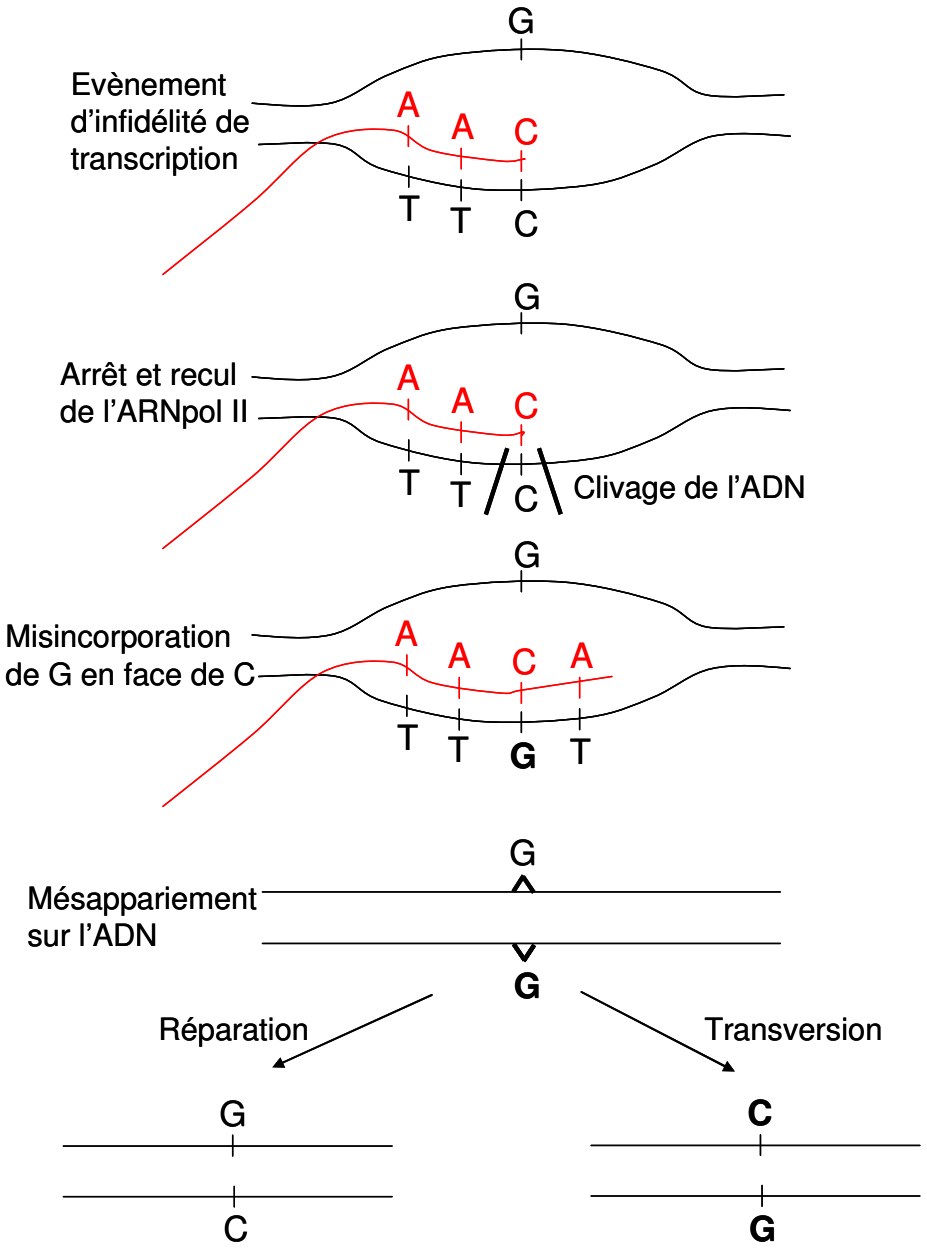


Figure 69 : Hypothèse d'un mécanisme générant les mutations somatiques.

Il est intéressant de noter que les mutations hypersomatiques qui caractérisent la maturation des immunoglobulines surviennent au cours de la transcription. La cascade d'événements présentée dans la figure ci-dessus n'est pas établie à ce jour, bien que chacune des étapes ait été décrite. Un des prochains objectifs du groupe sera de valider biologiquement l'existence de cette cascade d'évènements ; en effet, la recherche de mutations somatiques du cancer sera facilitée par la connaissance de l'origine de ces mutations.

En outre, tout comme les substitutions survenant au cours de la transcription, il est admis que la majorité des mutations somatiques survient en position 3 du codon, ce qui minimise l'impact codant, laissant supposer une moindre pression de sélection évolutive. Toutefois, si

d'une part on formule l'hypothèse selon laquelle certains évènements d'infidélité de transcription laissent une trace sur le génome sous forme de mutations somatiques et si d'autre part le variant protéique généré confère un avantage compétitif à la cellule germinale qui le produit, alors cette trace sera transmise à la descendance. Ceci ne représente qu'une hypothèse de travail mais présente l'avantage de proposer un mécanisme nouveau à l'évolution considérée aujourd'hui comme un phénomène aléatoire, mais qui pourrait un jour devenir dirigée. Cette hypothèse ainsi formulée serait en accord avec l'hypothèse de Bovery selon laquelle les propriétés d'une cellule cancéreuse lui seraient conférées par une accélération évolutive.

Ces avancées ont été rendues possibles par l'exploitation de plusieurs millions de séquences ESTs considérées comme des séquences de piètre qualité. Cette formidable source de données nous a permis de formuler l'hypothèse selon laquelle la transcription d'un même segment d'ADN pouvait générer différents variants, non fidèles à la séquence de l'ADN.
Si l'on ajoute à l'infidélité de transcription l'épissage alternatif, l'editing de l'ARN, le rôle des ARN interférents et une possible infidélité de traduction, il devient possible d'appréhender la complexité vraie de la cellule cancéreuse. Les stratégies diagnostiques et thérapeutiques devront tenir compte de cette réalité.

REFERENCES

1. Sirachy, J. An approach to the problem of heterogeneity of human tumour-cell populations. *Br J Cancer* **39**, 570-7 (1979).
2. Vindelov, L. L. et al. Clonal heterogeneity of small-cell anaplastic carcinoma of the lung demonstrated by flow-cytometric DNA analysis. *Cancer Res* **40**, 4295-300 (1980).
3. Merlo, L. M., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nat Rev Cancer* **6**, 924-35 (2006).
4. Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nat Med* **10**, 789-99 (2004).
5. Barbacid, M. ras genes. *Annu Rev Biochem* **56**, 779-827 (1987).
6. Mascaux, C. et al. The role of RAS oncogene in survival of patients with lung cancer: a systematic review of the literature with meta-analysis. *Br J Cancer* **92**, 131-9 (2005).
7. Hussain, S. P. et al. Increased p53 mutation load in noncancerous colon tissue from ulcerative colitis: a cancer-prone chronic inflammatory disease. *Cancer Res* **60**, 3333-7 (2000).
8. Payne, S. R. & Kemp, C. J. Tumor suppressor genetics. *Carcinogenesis* **26**, 2031-45 (2005).
9. Albertson, D. G., Collins, C., McCormick, F. & Gray, J. W. Chromosome aberrations in solid tumors. *Nat Genet* **34**, 369-76 (2003).
10. Ford, D. et al. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am J Hum Genet* **62**, 676-89 (1998).
11. Rebbeck, T. R. et al. Genetic heterogeneity in hereditary breast cancer: role of BRCA1 and BRCA2. *Am J Hum Genet* **59**, 547-53 (1996).
12. Olivier, M. et al. The IARC TP53 database: new online mutation analysis and recommendations to users. *Hum Mutat* **19**, 607-14 (2002).
13. Soussi, T., Kato, S., Levy, P. P. & Ishioka, C. Reassessment of the TP53 mutation database in human disease by data mining with a library of TP53 missense mutations. *Hum Mutat* **25**, 6-17 (2005).
14. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759-67 (1990).

15. Sjoblom, T. et al. The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268-74 (2006).
16. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23-8 (1976).
17. Loeb, L. A., Springgate, C. F. & Battula, N. Errors in DNA replication as a basis of malignant changes. *Cancer Res* **34**, 2311-21 (1974).
18. Bielas, J. H., Loeb, K. R., Rubin, B. P., True, L. D. & Loeb, L. A. Human cancers express a mutator phenotype. *Proc Natl Acad Sci U S A* **103**, 18238-42 (2006).
19. Loeb, L. A., Bielas, J. H. & Beckman, R. A. Cancers exhibit a mutator phenotype: clinical implications. *Cancer Res* **68**, 3551-7; discussion 3557 (2008).
20. Loeb, L. A., Loeb, K. R. & Anderson, J. P. Multiple mutations and cancer. *Proc Natl Acad Sci U S A* **100**, 776-81 (2003).
21. Bielas, J. H. & Loeb, L. A. Quantification of random genomic mutations. *Nat Methods* **2**, 285-90 (2005).
22. Guo, H. H., Choe, J. & Loeb, L. A. Protein tolerance to random amino acid change. *Proc Natl Acad Sci U S A* **101**, 9205-10 (2004).
23. Momparler, R. L. Cancer epigenetics. *Oncogene* **22**, 6479-83 (2003).
24. Baylin, S. B. et al. Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Hum Mol Genet* **10**, 687-92 (2001).
25. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41-5 (2000).
26. Berger, S. L. Histone modifications in transcriptional regulation. *Curr Opin Genet Dev* **12**, 142-8 (2002).
27. Yoder, J. A., Soman, N. S., Verdine, G. L. & Bestor, T. H. DNA (cytosine-5)-methyltransferases in mouse cells and tissues. Studies with a mechanism-based probe. *J Mol Biol* **270**, 385-95 (1997).
28. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J Mol Biol* **196**, 261-82 (1987).
29. Cottrell, S. E. Molecular diagnostic applications of DNA methylation technology. *Clin Biochem* **37**, 595-604 (2004).
30. Robertson, K. D. & Wolffe, A. P. DNA methylation in health and disease. *Nat Rev Genet* **1**, 11-9 (2000).
31. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A* **90**, 11995-9 (1993).

32. Momparler, R. L. & Bovenzi, V. DNA methylation and cancer. *J Cell Physiol* **183**, 145-54 (2000).
33. Luczak, M. W. & Jagodzinski, P. P. The role of DNA methylation in cancer development. *Folia Histochem Cytobiol* **44**, 143-54 (2006).
34. Girault, I., Tozlu, S., Lidereau, R. & Bieche, I. Expression analysis of DNA methyltransferases 1, 3A, and 3B in sporadic breast carcinomas. *Clin Cancer Res* **9**, 4415-22 (2003).
35. Singal, R., Das, P. M., Manoharan, M., Reis, I. M. & Schlesselman, J. J. Polymorphisms in the DNA methyltransferase 3b gene and prostate cancer risk. *Oncol Rep* **14**, 569-73 (2005).
36. Schneider-Stock, R. et al. 5-Aza-cytidine is a potent inhibitor of DNA methyltransferase 3a and induces apoptosis in HCT-116 colon cancer cells via Gadd45- and p53-dependent mechanisms. *J Pharmacol Exp Ther* **312**, 525-36 (2005).
37. Xiong, Y. et al. Opposite alterations of DNA methyltransferase gene expression in endometrioid and serous endometrial cancers. *Gynecol Oncol* **96**, 601-9 (2005).
38. Futreal, P. A. et al. A census of human cancer genes. *Nat Rev Cancer* **4**, 177-83 (2004).
39. Forbes, S. et al. Cosmic 2005. *Br J Cancer* **94**, 318-22 (2006).
40. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-70 (1995).
41. van de Vijver, M. J. et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**, 1999-2009 (2002).
42. van 't Veer, L. J. et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-6 (2002).
43. Bueno-de-Mesquita, J. M. et al. Validation of 70-gene prognosis signature in node-negative breast cancer. *Breast Cancer Res Treat* (2008).
44. Buyse, M. et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst* **98**, 1183-92 (2006).
45. Wang, Y. et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671-9 (2005).
46. Foekens, J. A. et al. Multicenter validation of a gene expression-based prognostic signature in lymph node-negative primary breast cancer. *J Clin Oncol* **24**, 1665-71 (2006).

47. Haibe-Kains, B. et al. Comparison of prognostic gene expression signatures for breast cancer. *BMC Genomics* **9**, 394 (2008).
48. Rosenwald, A. et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med* **346**, 1937-47 (2002).
49. Shipp, M. A. et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* **8**, 68-74 (2002).
50. Tan, P. K. et al. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* **31**, 5676-84 (2003).
51. Marshall, E. Getting the noise out of gene arrays. *Science* **306**, 630-1 (2004).
52. Draghici, S., Khatri, P., Eklund, A. C. & Szallasi, Z. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet* **22**, 101-9 (2006).
53. Eklund, A. C. & Szallasi, Z. Correction of technical bias in clinical microarray data improves concordance with known biological information. *Genome Biol* **9**, R26 (2008).
54. Halgren, R. G., Fielden, M. R., Fong, C. J. & Zacharewski, T. R. Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones. *Nucleic Acids Res* **29**, 582-8 (2001).
55. Taylor, E. et al. Sequence verification as quality-control step for production of cDNA microarrays. *Biotechniques* **31**, 62-5 (2001).
56. Harbig, J., Sprinkle, R. & Enkemann, S. A. A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. *Nucleic Acids Res* **33**, e31 (2005).
57. Mecham, B. H. et al. Increased measurement accuracy for sequence-verified microarray probes. *Physiol Genomics* **18**, 308-15 (2004).
58. Barrett, J. C. & Kawasaki, E. S. Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. *Drug Discov Today* **8**, 134-41 (2003).
59. Zhang, J., Finney, R. P., Clifford, R. J., Derr, L. K. & Buetow, K. H. Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach. *Genomics* **85**, 297-308 (2005).
60. Kane, M. D. et al. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res* **28**, 4552-7 (2000).
61. Holland, M. J. Transcript abundance in yeast varies over six orders of magnitude. *J Biol Chem* **277**, 14363-6 (2002).

62. Czechowski, T., Bari, R. P., Stitt, M., Scheible, W. R. & Udvardi, M. K. Real-time RT-PCR profiling of over 1400 Arabidopsis transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *Plant J* **38**, 366-79 (2004).
63. Bakay, M. et al. Sources of variability and effect of experimental approach on expression profiling data interpretation. *BMC Bioinformatics* **3**, 4 (2002).
64. Bammler, T. et al. Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods* **2**, 351-6 (2005).
65. Houseley, J., LaCava, J. & Tollervey, D. RNA-quality control by the exosome. *Nat Rev Mol Cell Biol* **7**, 529-39 (2006).
66. Isken, O. & Maquat, L. E. Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes Dev* **21**, 1833-56 (2007).
67. Stalder, L. & Muhlemann, O. The meaning of nonsense. *Trends Cell Biol* **18**, 315-21 (2008).
68. Le Hir, H., Gatfield, D., Izaurralde, E. & Moore, M. J. The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *Embo J* **20**, 4987-97 (2001).
69. Nagy, E. & Maquat, L. E. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem Sci* **23**, 198-9 (1998).
70. Chang, Y. F., Imam, J. S. & Wilkinson, M. F. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* **76**, 51-74 (2007).
71. Miriami, E., Motro, U., Sperling, J. & Sperling, R. Conservation of an open-reading frame as an element affecting 5' splice site selection. *J Struct Biol* **140**, 116-22 (2002).
72. Kim, Y. K. et al. Staufen1 regulates diverse classes of mammalian transcripts. *Embo J* **26**, 2670-81 (2007).
73. Kim, Y. K., Furic, L., Desgroseillers, L. & Maquat, L. E. Mammalian Staufen1 recruits Upf1 to specific mRNA 3'UTRs so as to elicit mRNA decay. *Cell* **120**, 195-208 (2005).
74. Wagner, E. & Lykke-Andersen, J. mRNA surveillance: the perfect persist. *J Cell Sci* **115**, 3033-8 (2002).
75. Jeon, C. & Agarwal, K. Fidelity of RNA polymerase II transcription controlled by elongation factor TFIIS. *Proc Natl Acad Sci U S A* **93**, 13677-82 (1996).
76. Thomas, M. J., Platas, A. A. & Hawley, D. K. Transcriptional fidelity and proofreading by RNA polymerase II. *Cell* **93**, 627-37 (1998).

77. Moussard C, M. C. *Biologie moléculaire. Biochimie des communications cellulaires* (ed. De Boeck Université) (2005).
78. Pomerantz, R. T., Temiakov, D., Anikin, M., Vassylyev, D. G. & McAllister, W. T. A mechanism of nucleotide misincorporation during transcription due to template-strand misalignment. *Mol Cell* **24**, 245-55 (2006).
79. Kashkina, E. et al. Template misalignment in multisubunit RNA polymerases and transcription fidelity. *Mol Cell* **24**, 257-66 (2006).
80. Evans, D. A., van der Kleij, A. A., Sonnemans, M. A., Burbach, J. P. & van Leeuwen, F. W. Frameshift mutations at two hotspots in vasopressin transcripts in post-mitotic neurons. *Proc Natl Acad Sci U S A* **91**, 6059-63 (1994).
81. van Leeuwen, F. W. et al. Molecular misreading: a new type of transcript mutation expressed during aging. *Neurobiol Aging* **21**, 879-91 (2000).
82. van Leeuwen, F. W., Kros, J. M., Kamphorst, W., van Schravendijk, C. & de Vos, R. A. Molecular misreading: the occurrence of frameshift proteins in different diseases. *Biochem Soc Trans* **34**, 738-42 (2006).
83. Linton, M. F., Pierotti, V. & Young, S. G. Reading-frame restoration with an apolipoprotein B gene frameshift mutation. *Proc Natl Acad Sci U S A* **89**, 11431-5 (1992).
84. Young, M. et al. Partial correction of a severe molecular defect in hemophilia A, because of errors during expression of the factor VIII gene. *Am J Hum Genet* **60**, 565-73 (1997).
85. Ba, Y. et al. Transcriptional slippage of p53 gene enhanced by cellular damage in rat liver: monitoring the slippage by a yeast functional assay. *Mutat Res* **447**, 209-20 (2000).
86. Benson, K. F., Person, R. E., Li, F. Q., Williams, K. & Horwitz, M. Paradoxical homozygous expression from heterozygotes and heterozygous expression from homozygotes as a consequence of transcriptional infidelity through a polyadenine tract in the AP3B1 gene responsible for canine cyclic neutropenia. *Nucleic Acids Res* **32**, 6327-33 (2004).
87. Adams, M. D. et al. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651-6 (1991).
88. Okubo, K. et al. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat Genet* **2**, 173-9 (1992).

89. Aaronson, J. S. et al. Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. *Genome Res* **6**, 829-45 (1996).

90. Liang, F. et al. An optimized protocol for analysis of EST sequences. *Nucleic Acids Res* **28**, 3657-65 (2000).

91. Adams, M. D., Kerlavage, A. R., Fields, C. & Venter, J. C. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nat Genet* **4**, 256-67 (1993).

92. Adams, M. D., Soares, M. B., Kerlavage, A. R., Fields, C. & Venter, J. C. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat Genet* **4**, 373-80 (1993).

93. Adams, M. D. et al. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377**, 3-174 (1995).

94. Khan, A. S. et al. Single pass sequencing and physical and genetic mapping of human brain cDNAs. *Nat Genet* **2**, 180-5 (1992).

95. McCombie, W. R. et al. Caenorhabditis elegans expressed sequence tags identify gene families and potential disease gene homologues. *Nat Genet* **1**, 124-31 (1992).

96. Matsubara, K. & Okubo, K. Identification of new genes by systematic analysis of cDNAs and database construction. *Curr Opin Biotechnol* **4**, 672-7 (1993).

97. Hotz-Wagenblatt, A. et al. ESTAnnotator: A tool for high throughput EST annotation. *Nucleic Acids Res* **31**, 3716-9 (2003).

98. Jiang, J. & Jacob, H. J. EbEST: an automated tool using expressed sequence tags to delineate gene structure. *Genome Res* **8**, 268-75 (1998).

99. Kan, Z., Rouchka, E. C., Gish, W. R. & States, D. J. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res* **11**, 889-900 (2001).

100. Brett, D. et al. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett* **474**, 83-6 (2000).

101. Modrek, B., Resch, A., Grasso, C. & Lee, C. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res* **29**, 2850-9 (2001).

102. Kan, Z., Castle, J., Johnson, J. M. & Tsinoremas, N. F. Detection of novel splice forms in human and mouse using cross-species approach. *Pac Symp Biocomput*, 42-53 (2004).

103. Dieterich, C., Wang, H., Rateitschak, K., Luz, H. & Vingron, M. CORG: a database for COmparative Regulatory Genomics. *Nucleic Acids Res* **31**, 55-7 (2003).
104. Mach, V. PRESTA: associating promoter sequences with information on gene expression. *Genome Biol* **3**, research0050 (2002).
105. Beaudoing, E. & Gautheret, D. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Res* **11**, 1520-6 (2001).
106. Gautheret, D., Poirot, O., Lopez, F., Audic, S. & Claverie, J. M. Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res* **8**, 524-30 (1998).
107. Yan, J. & Marr, T. G. Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Res* **15**, 369-75 (2005).
108. Bonaldo, M. F., Lennon, G. & Soares, M. B. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res* **6**, 791-806 (1996).
109. Schmitt, A. O. et al. Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Res* **27**, 4251-60 (1999).
110. Aouacheria, A., Navratil, V., Barthelaix, A., Mouchiroud, D. & Gautier, C. Bioinformatic screening of human ESTs for differentially expressed genes in normal and tumor tissues. *BMC Genomics* **7**, 94 (2006).
111. Buetow, K. H., Edmonson, M. N. & Cassidy, A. B. Reliable identification of large numbers of candidate SNPs from public EST data. *Nat Genet* **21**, 323-5 (1999).
112. Picoult-Newberg, L. et al. Mining SNPs from EST databases. *Genome Res* **9**, 167-74 (1999).
113. Marth, G. T. et al. A general approach to single-nucleotide polymorphism discovery. *Nat Genet* **23**, 452-6 (1999).
114. Useche, F. J., Gao, G., Harafey, M. & Rafalski, A. High-throughput identification, database storage and analysis of SNPs in EST sequences. *Genome Inform* **12**, 194-203 (2001).
115. Aouacheria, A. et al. In silico whole-genome screening for cancer-related single-nucleotide polymorphisms located in human mRNA untranslated regions. *BMC Genomics* **8**, 2 (2007).
116. Datson, N. A. et al. Development of the first marmoset-specific DNA microarray (EUMAMA): a new genetic tool for large-scale expression profiling in a non-human primate. *BMC Genomics* **8**, 190 (2007).

117. Forment, J. et al. EST2uni: an open, parallel tool for automated EST analysis and database creation, with a data mining web interface and microarray expression data integration. *BMC Bioinformatics* **9**, 5 (2008).
118. Mao, C., Cushman, J. C., May, G. D. & Weller, J. W. ESTAP--an automated system for the analysis of EST data. *Bioinformatics* **19**, 1720-2 (2003).
119. Parkinson, J. et al. PartiGene--constructing partial genomes. *Bioinformatics* **20**, 1398-404 (2004).
120. Nagaraj, S. H., Deshpande, N., Gasser, R. B. & Ranganathan, S. ESTExplorer: an expressed sequence tag (EST) assembly and annotation platform. *Nucleic Acids Res* **35**, W143-7 (2007).
121. Waegele, B., Schmidt, T., Mewes, H. W. & Ruepp, A. OREST: the online resource for EST analysis. *Nucleic Acids Res* **36**, W140-4 (2008).
122. Lottaz, C., Iseli, C., Jongeneel, C. V. & Bucher, P. Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics* **19 Suppl 2**, II103-II112 (2003).
123. Collins, F. S. et al. New goals for the U.S. Human Genome Project: 1998-2003. *Science* **282**, 682-9 (1998).
124. Nagaraj, S. H., Gasser, R. B. & Ranganathan, S. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief Bioinform* **8**, 6-21 (2007).
125. Hillier, L. D. et al. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res* **6**, 807-28 (1996).
126. Lee, B. & Shin, G. CleanEST: a database of cleansed EST libraries. *Nucleic Acids Res* **37**, D686-9 (2009).
127. Boguski, M. S., Lowe, T. M. & Tolstoshev, C. M. dbEST--database for "expressed sequence tags". *Nat Genet* **4**, 332-3 (1993).
128. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res* **37**, D26-31 (2009).
129. Wheeler, D. L. et al. Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res* **32 Database issue**, D35-40 (2004).
130. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186-94 (1998).
131. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Res* **9**, 868-77 (1999).

132. Miller, R. T. et al. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res* **9**, 1143-55 (1999).

133. Lee, Y. et al. The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res* **33**, D71-4 (2005).

134. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-5 (2007).

135. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).

136. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**, 203-14 (2000).

137. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-11 (2001).

138. Dalmasso, C., Broet, P. Procédures d'estimation du false discovery rate basées sur la distribution des degrés de signification. *Journal de la Société Française de Statistiques* **146** (2005).

139. Armache, K. J., Kettenberger, H. & Cramer, P. The dynamic machinery of mRNA elongation. *Curr Opin Struct Biol* **15**, 197-203 (2005).

140. Olivier, M., Hussain, S. P., Caron de Fromentel, C., Hainaut, P. & Harris, C. C. TP53 mutation spectra and load: a tool for generating hypotheses on the etiology of cancer. *IARC Sci Publ*, 247-70 (2004).

141. Costello, J. F. et al. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. *Nat Genet* **24**, 132-8 (2000).

142. Holliday, R. & Grigg, G. W. DNA methylation and mutation. *Mutat Res* **285**, 61-7 (1993).

143. Walsh, C. P. & Xu, G. L. Cytosine methylation and DNA repair. *Curr Top Microbiol Immunol* **301**, 283-315 (2006).

144. Loechler, E. L., Green, C. L. & Essigmann, J. M. In vivo mutagenesis by O6-methylguanine built into a unique site in a viral genome. *Proc Natl Acad Sci U S A* **81**, 6271-5 (1984).

145. Green, C. L., Loechler, E. L., Fowler, K. W. & Essigmann, J. M. Construction and characterization of extrachromosomal probes for mutagenesis by carcinogens: site-

specific incorporation of O6-methylguanine into viral and plasmid genomes. *Proc Natl Acad Sci U S A* **81**, 13-7 (1984).

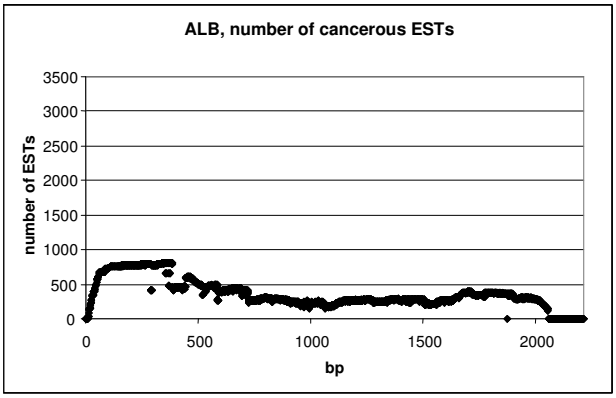
146. Warren, J. J., Forsberg, L. J. & Beese, L. S. The structural basis for the mutagenicity of O(6)-methyl-guanine lesions. *Proc Natl Acad Sci U S A* **103**, 19701-6 (2006).
147. Strachan, T. R. *Human molecular genetics* (Bios scientific, Oxford, 1999).
148. Amalric, F. *Analyse économique des coûts du cancer en France* (ed. Cancer, I. N. d.) (2007).
149. Collectif, F. n. d. c. d. l. c. l. c. F., A. Depierre, Société de pneumologie de langue française. *Cancers bronchopulmonaires non à petites cellules* (John Libbey Eurotext, 2002).
150. Sayers, E. W. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **37**, D5-15 (2009).

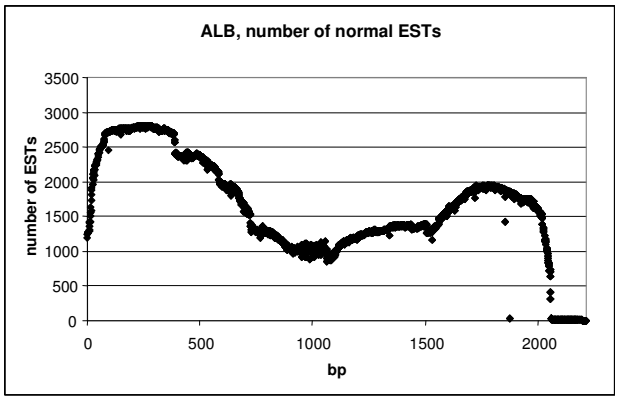
ANNEXE

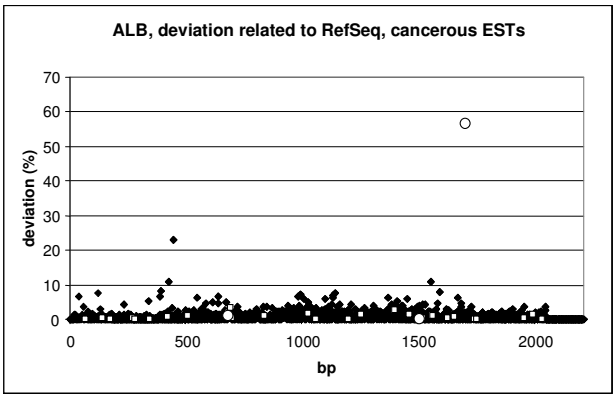
Résultats de l'étude préliminaire portant sur 16 transcrits.

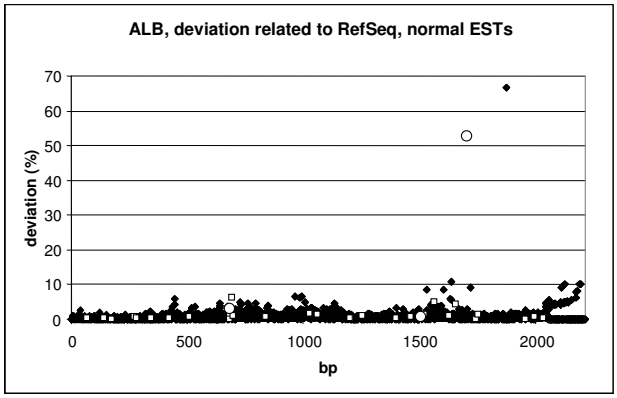
Pour chaque transcrit sont représentés le nombre d'ESTs (haut) à chaque position de l'ARNm ainsi que le pourcentage de déviations par rapport à la séquence de référence (bas) à chaque position de l'ARNm. Les résultats portant sur les ESTs isues de tissus cancéreux sont à gauche et issues de tissus normaux à droite.

Les SNPs validés biologiquement sont représentés par des ronds (○) et les autres par des carrés (□).

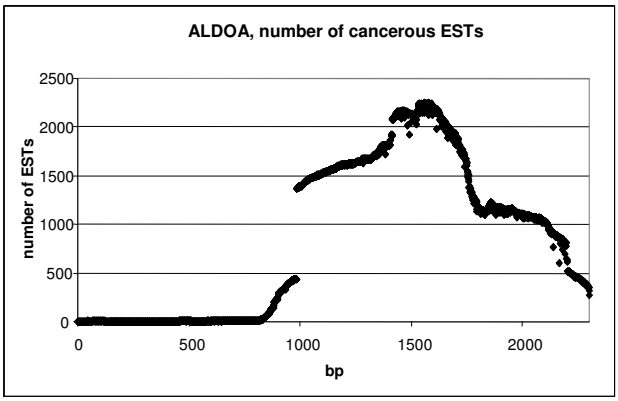


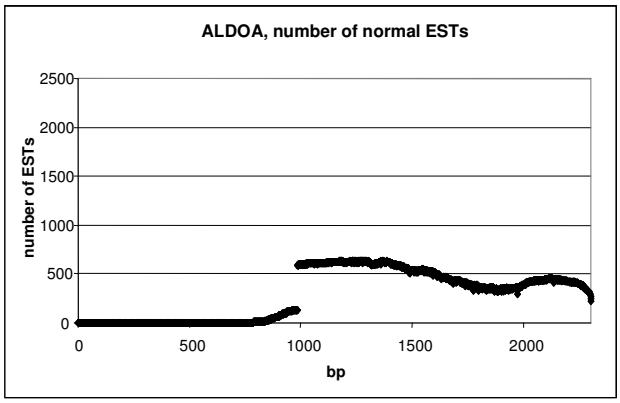


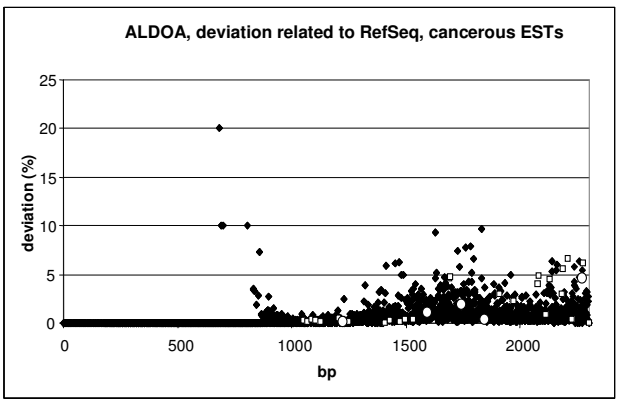


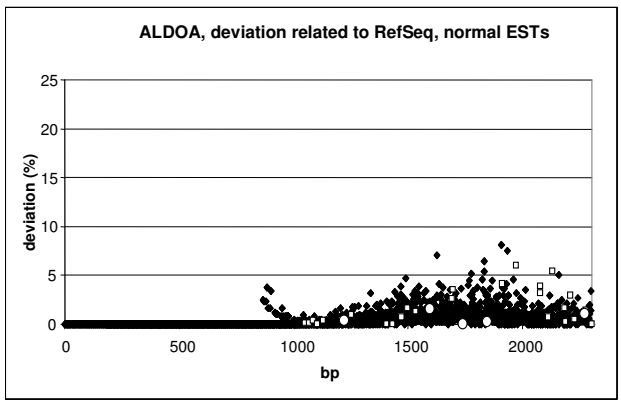


Albumin

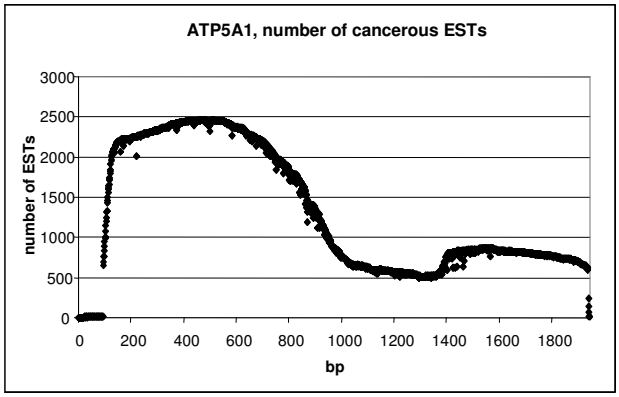


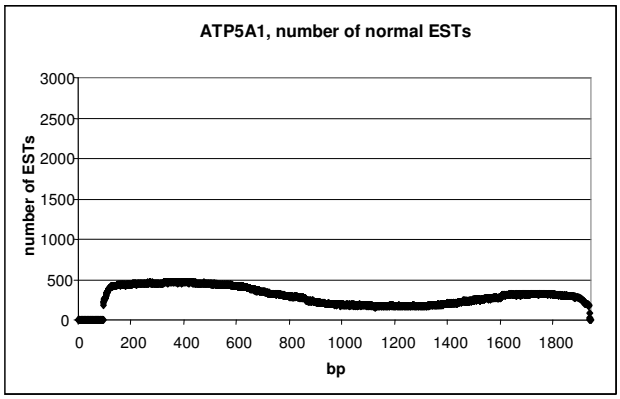


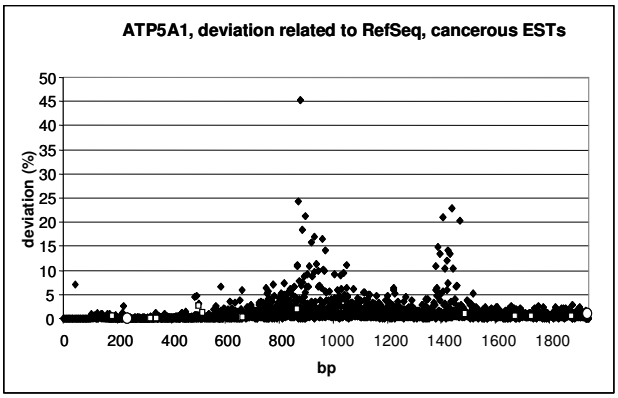


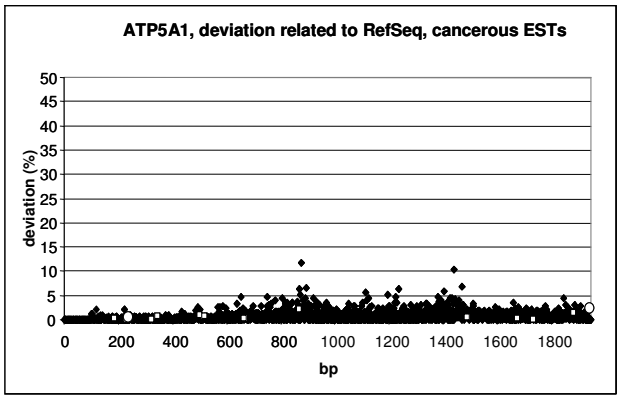


Aldolase A, fructose-bisphosphate

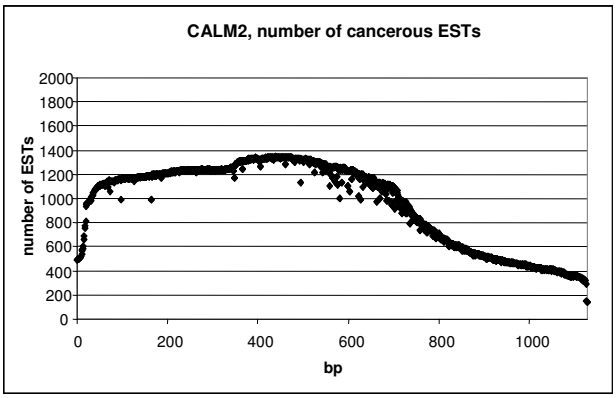


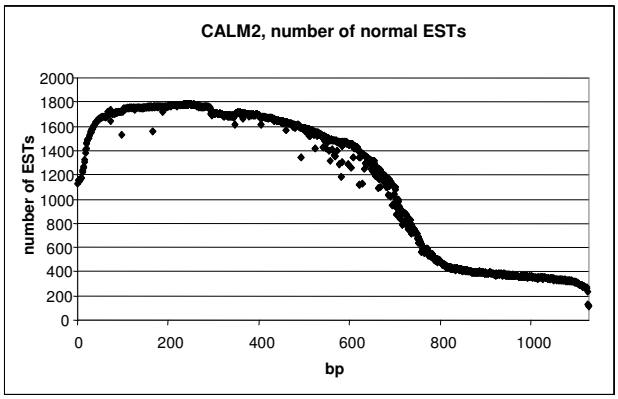


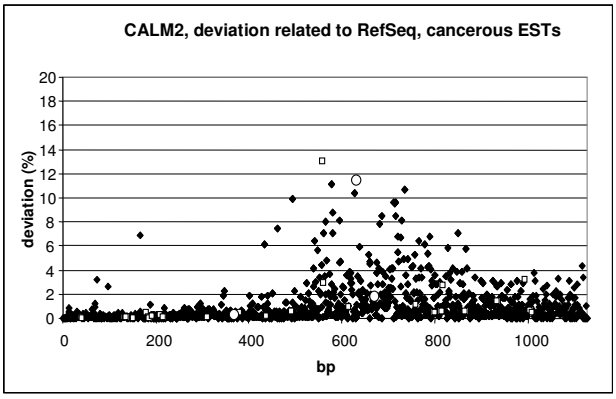


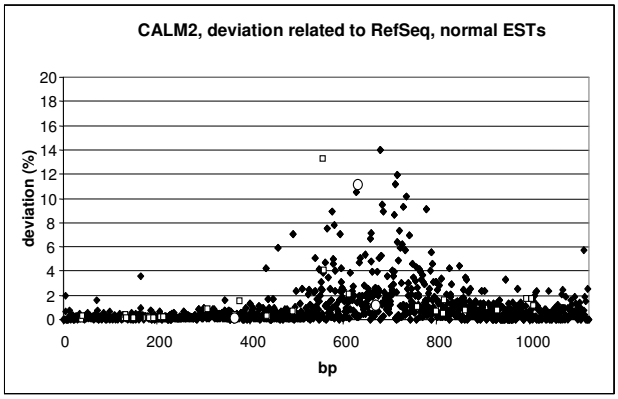


ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit 1, cardiac muscle

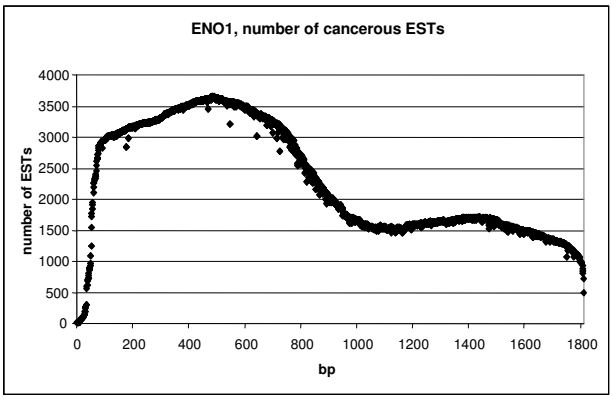


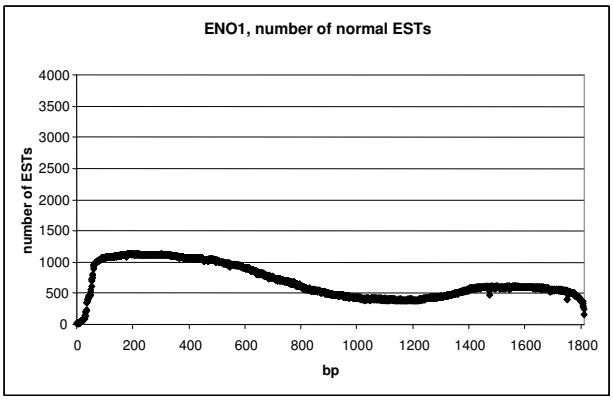


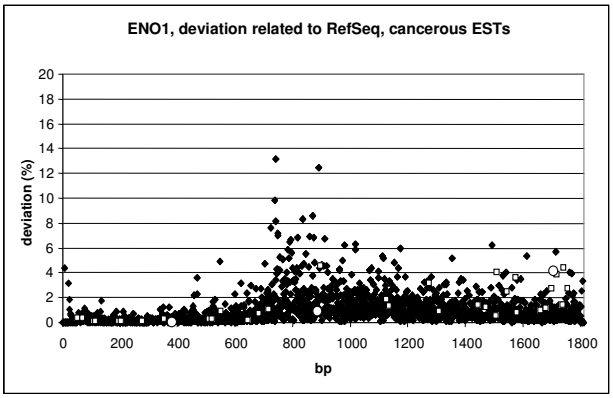


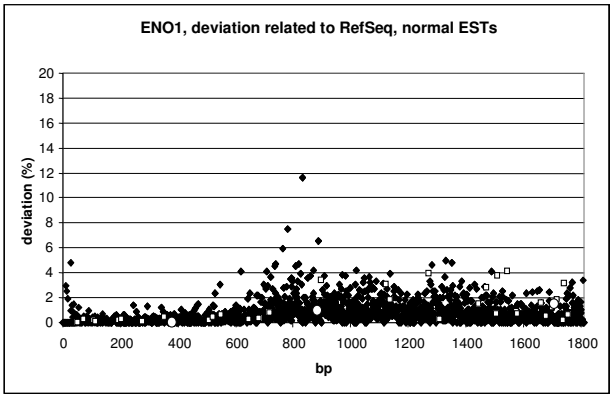


Calmodulin 2 (phosphorylase kinase, delta)

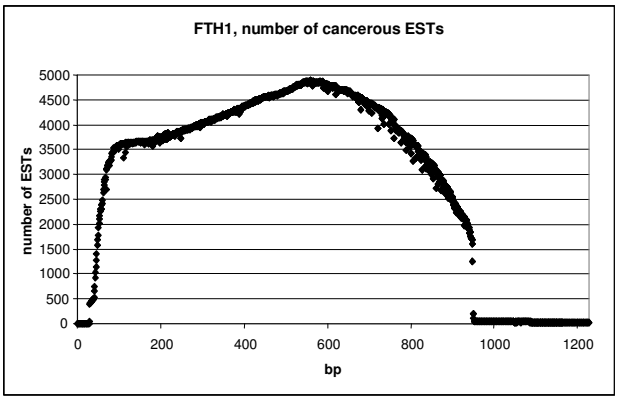


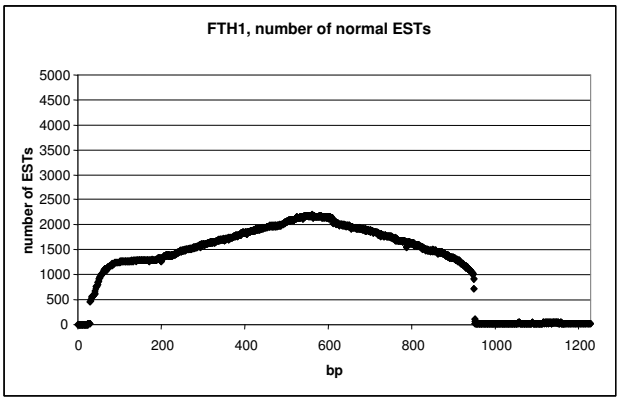


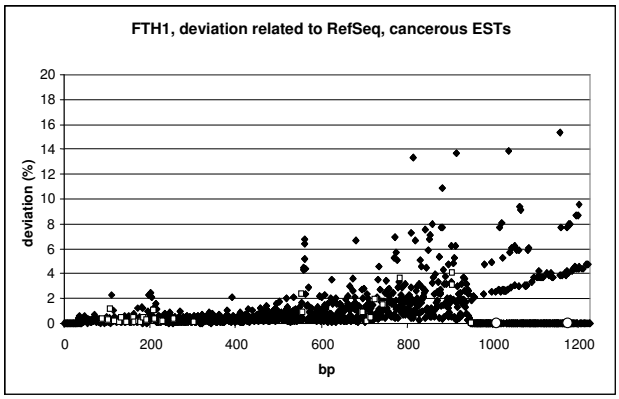


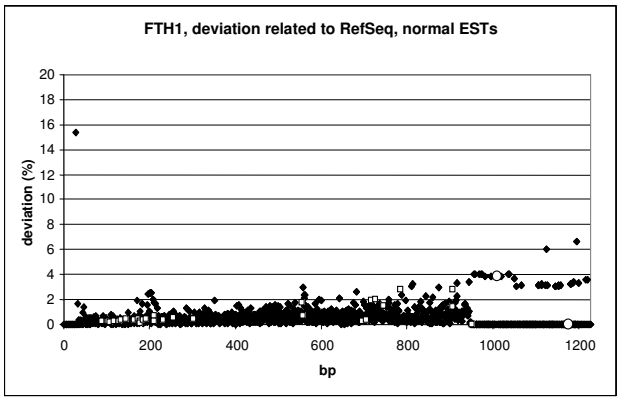


Enolase 1, (alpha)

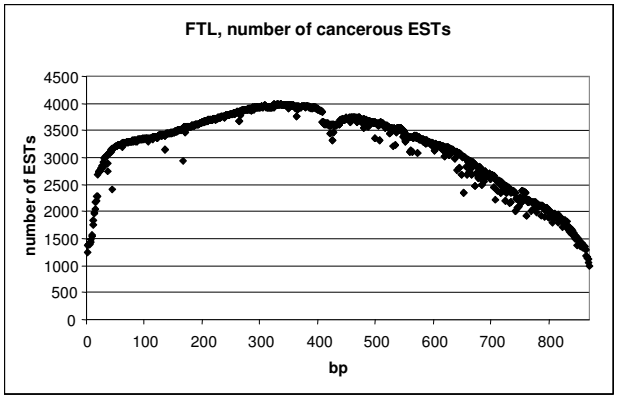


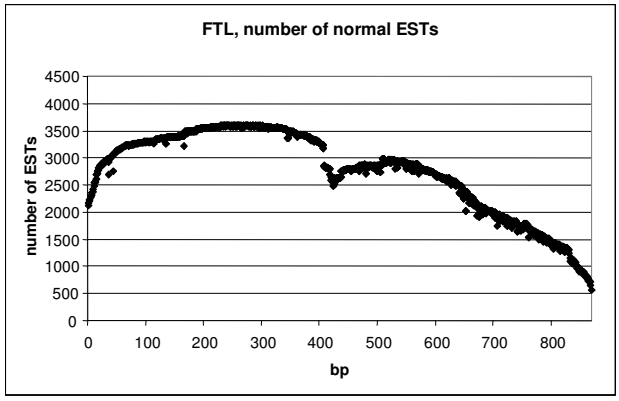


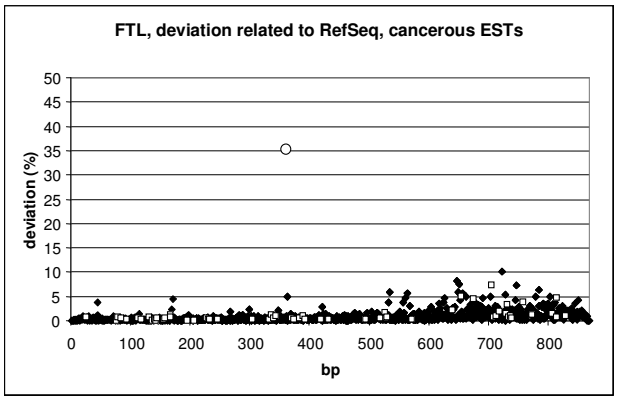


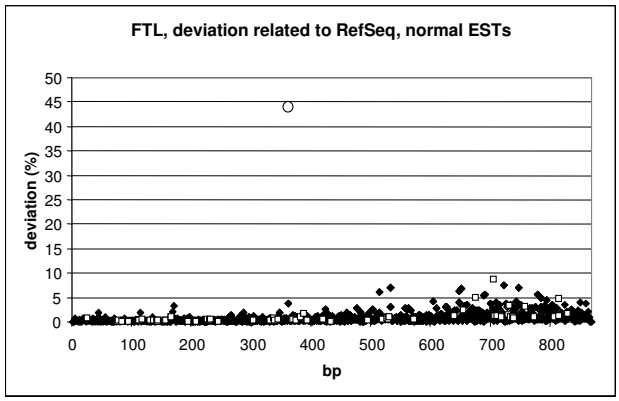


Ferritin, heavy polypeptide 1

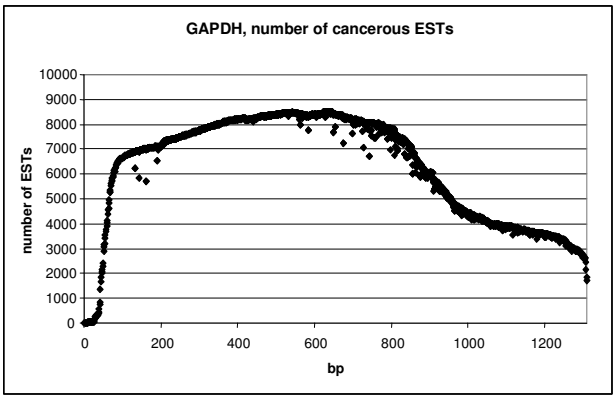


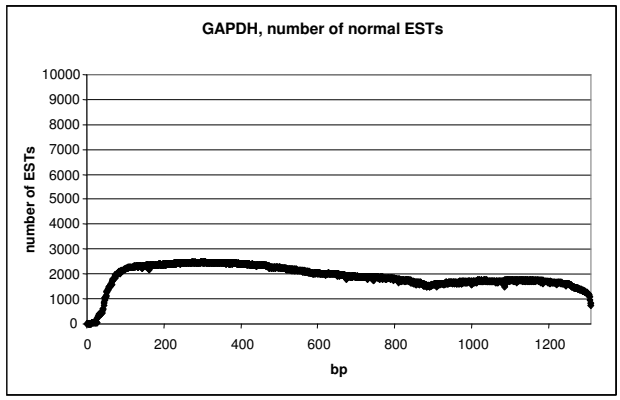


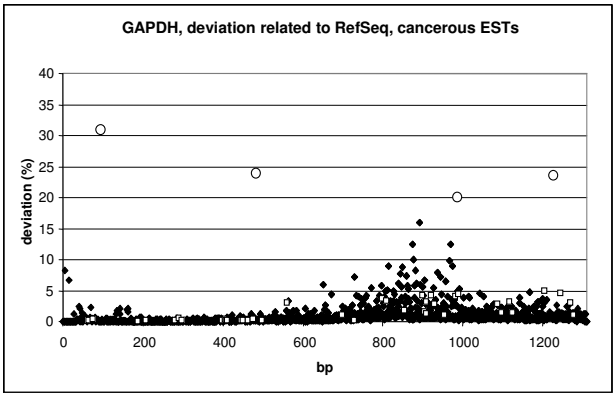


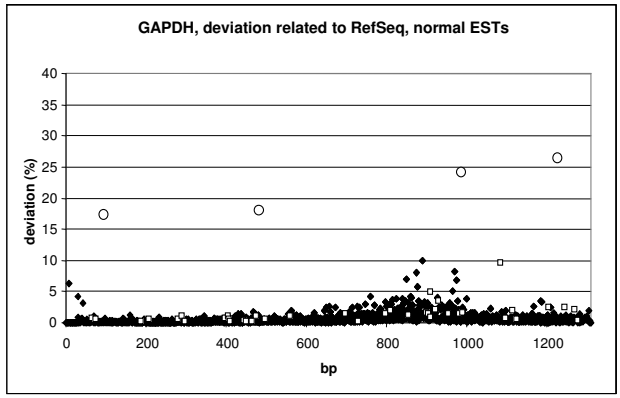


Ferritin, light polypeptide

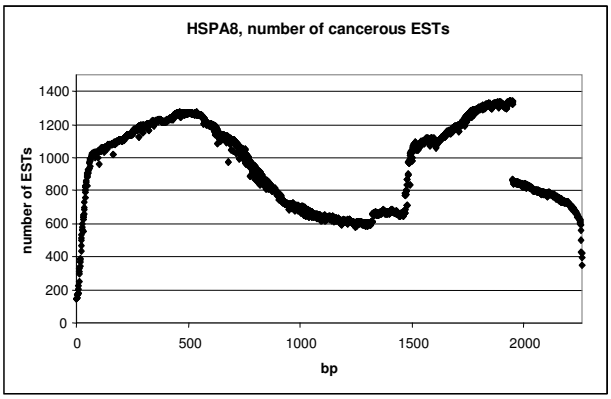


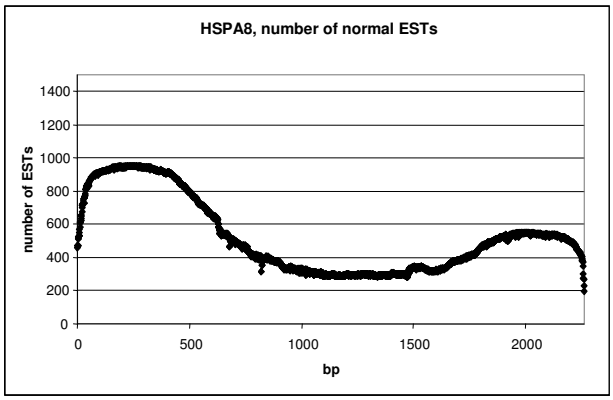


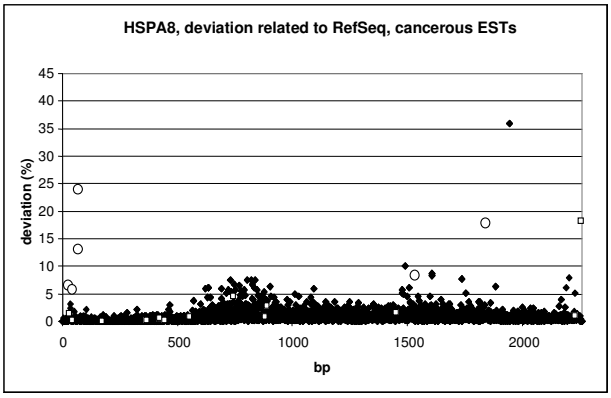


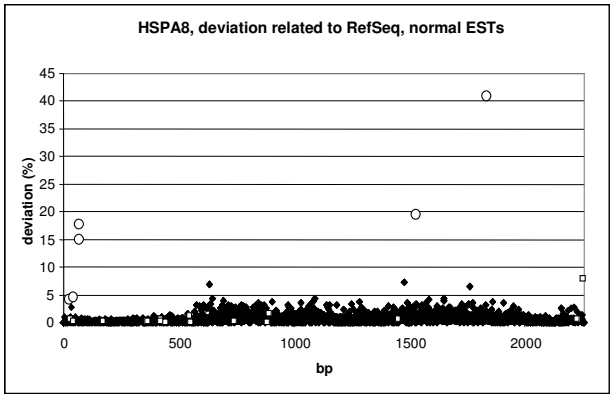


Glyceraldehyde-3-phosphate dehydrogenase

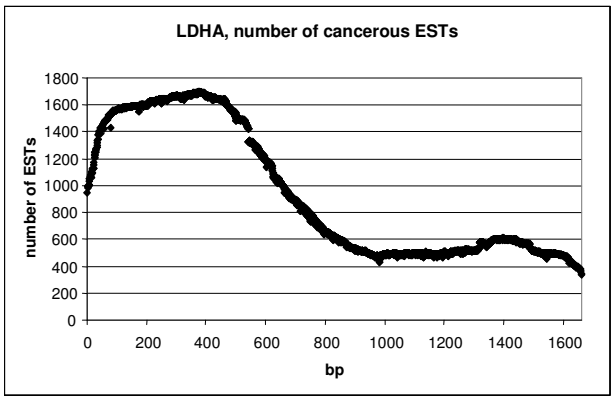


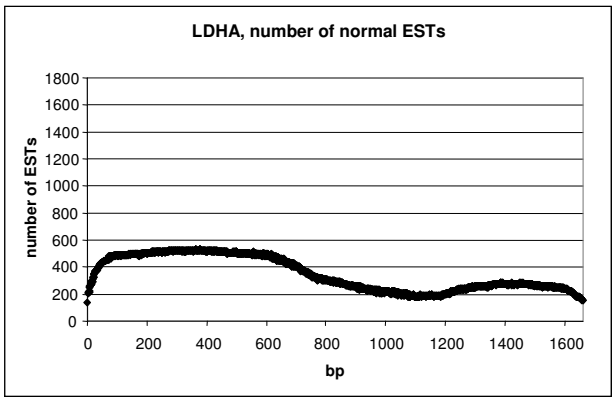


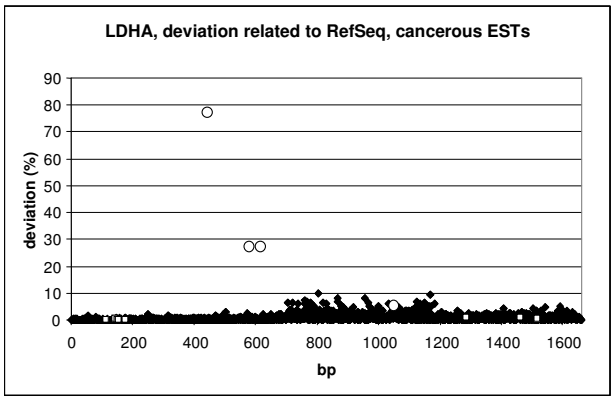


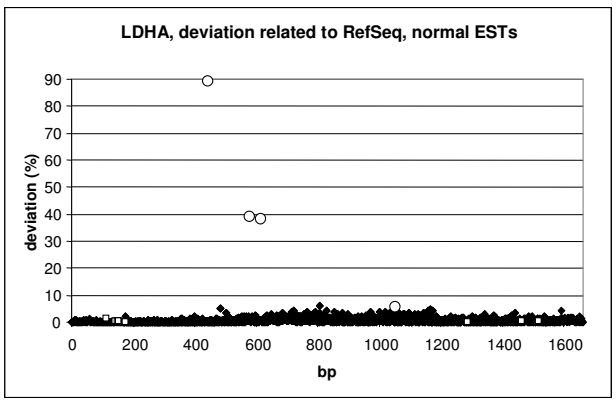


Heat shock 70kDa protein 8

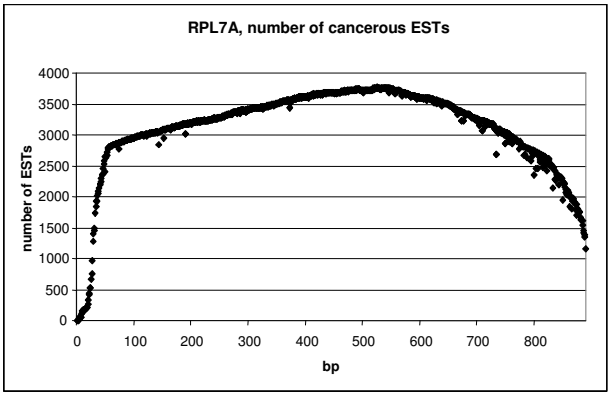


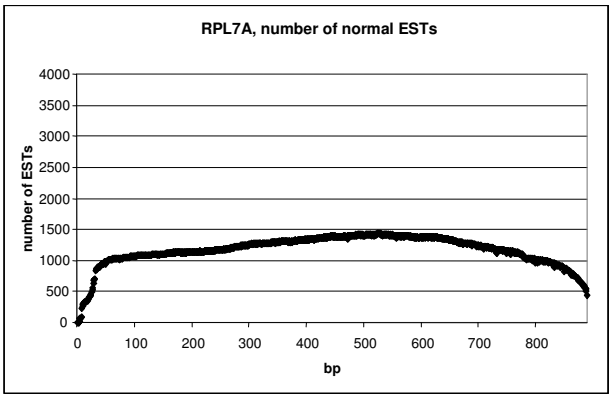


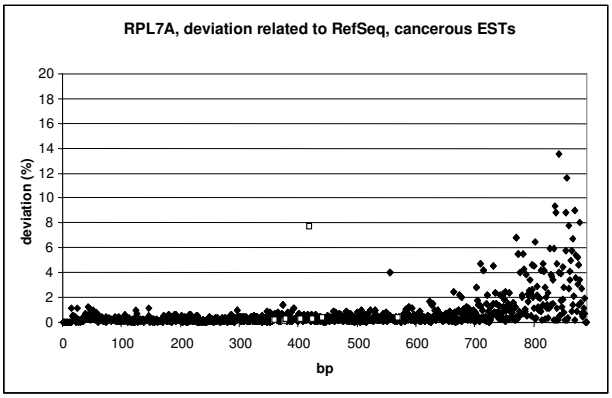


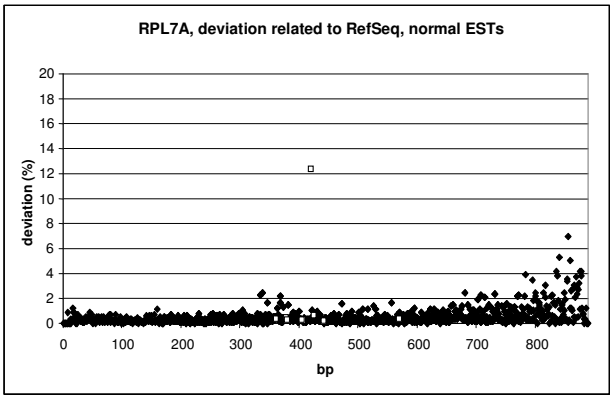


Lactate dehydrogenase A

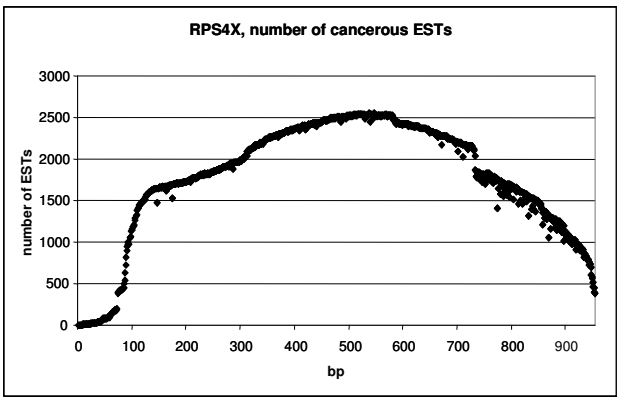


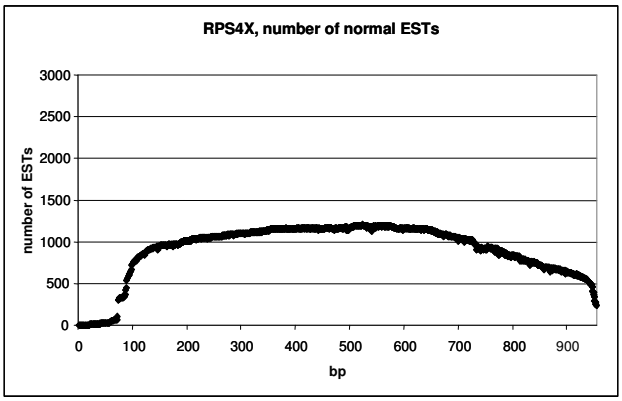


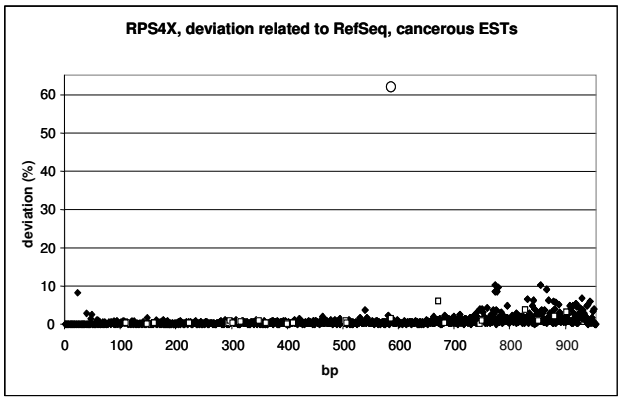


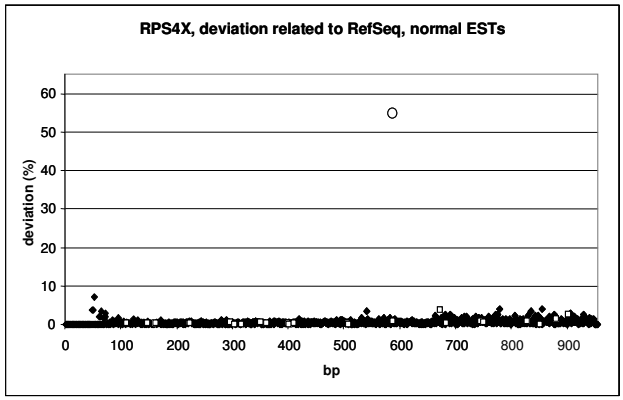


Ribosomal protein L7a

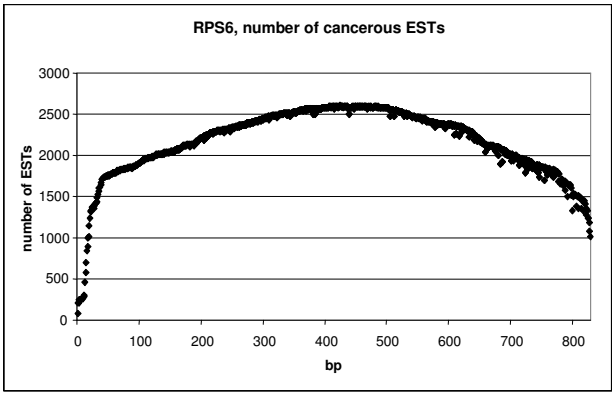


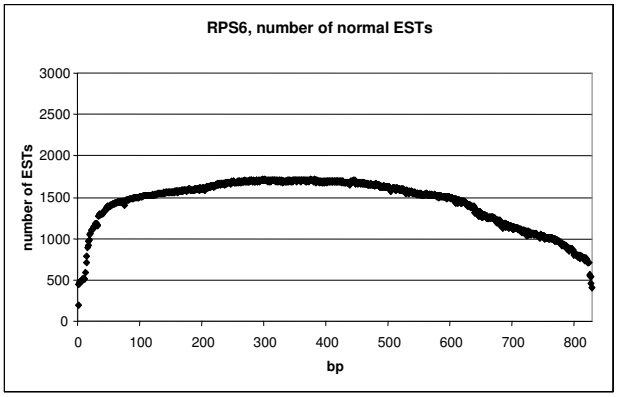


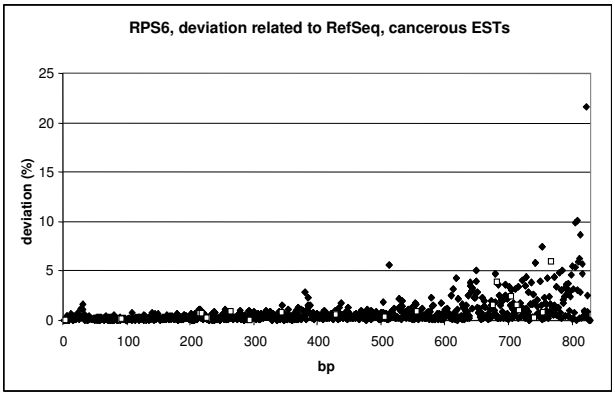


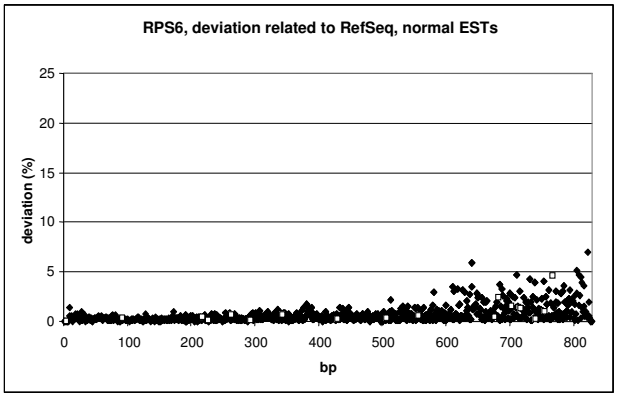


Ribosomal protein S4, X-linked

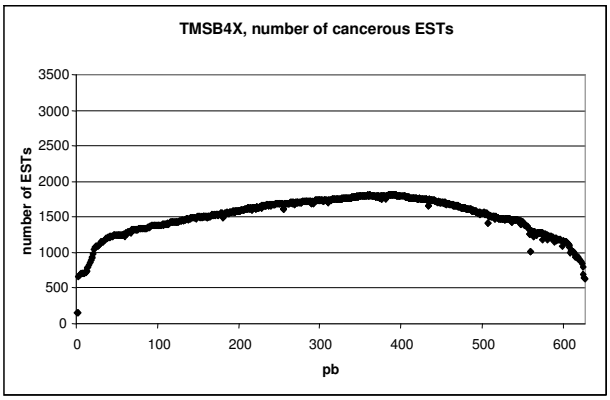


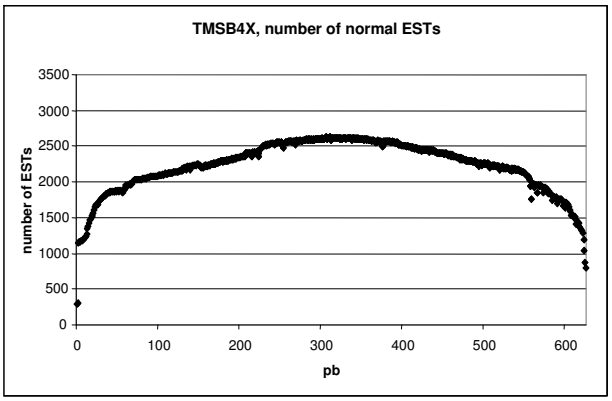


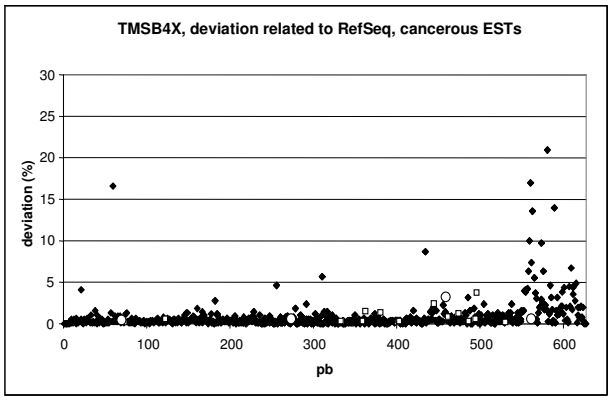


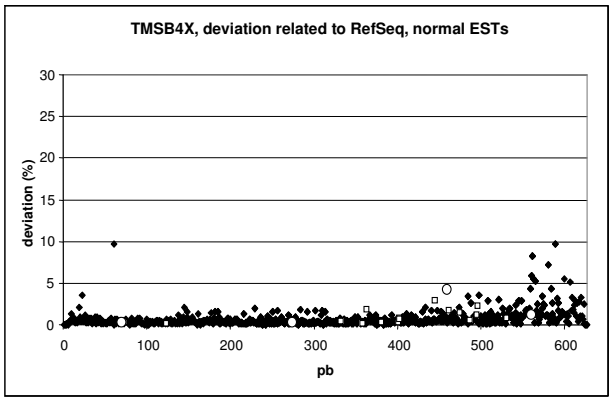


Ribosomal protein S6

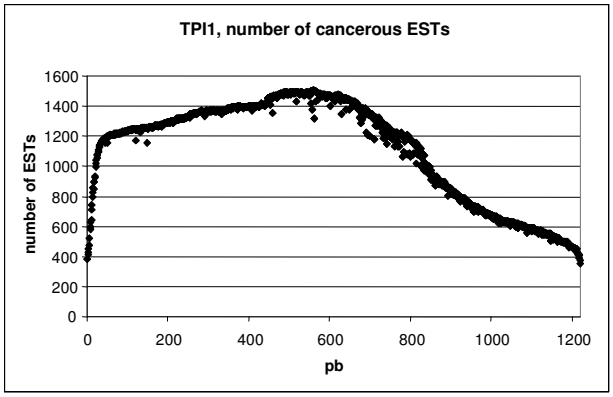


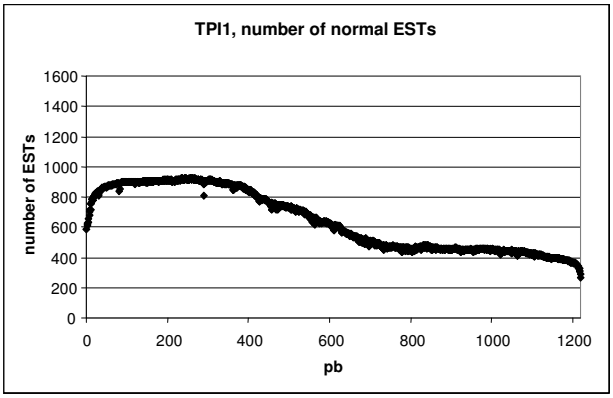


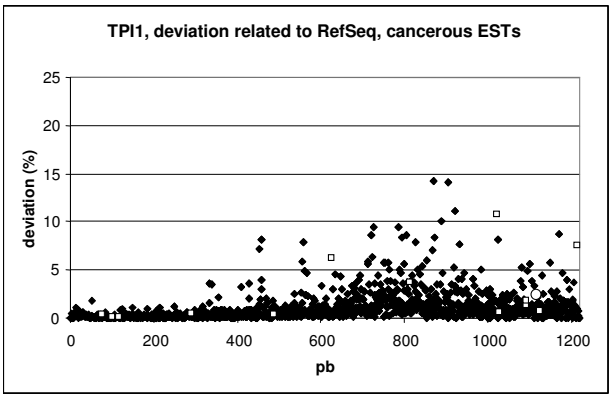


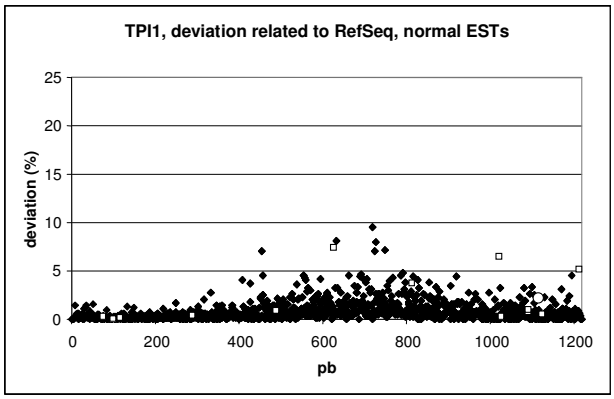


Thymosin, beta 4, X-linked

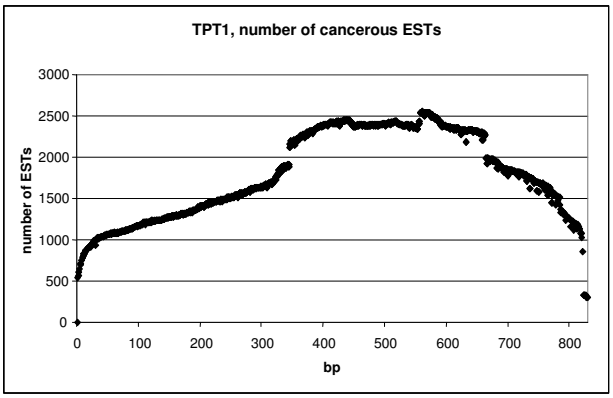


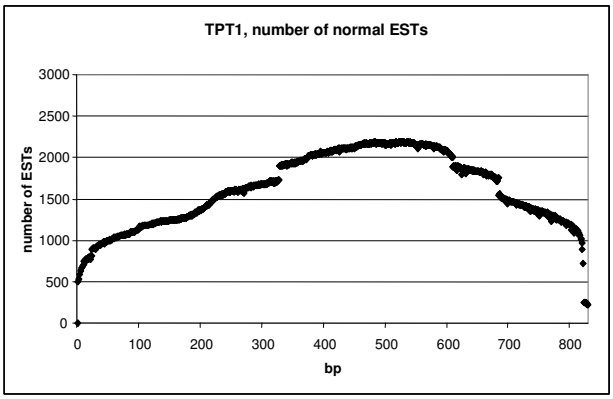


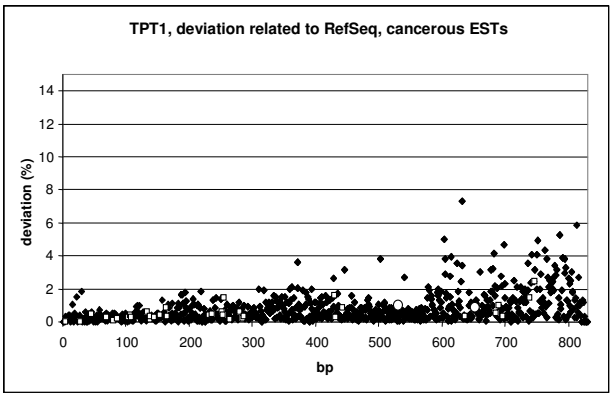


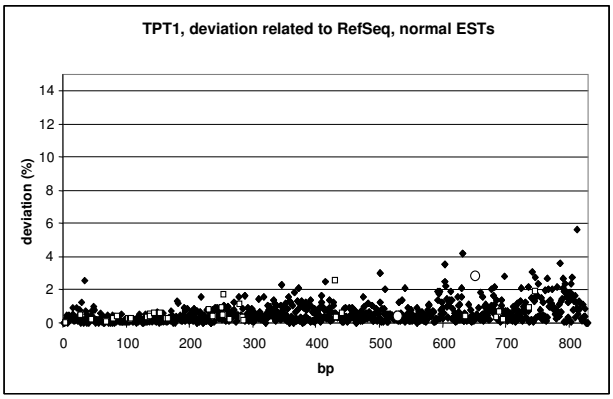


Triosephosphate isomerase 1









Tumor protein, translationally-controlled 1

INSTITUT NATIONAL
POLYTECHNIQUE
DE LORRAINE

AUTORISATION DE SOUTENANCE DE THESE
DU DOCTORAT DE L'INSTITUT NATIONAL
POLYTECHNIQUE DE LORRAINE

o0o

VU LES RAPPORTS ETABLIS PAR :
Monsieur François AMALRIC, Professeur, IPBS, Toulouse

Monsieur Jérôme CHAILLOUX, Directeur de Recherche, ERCIM

Le Président de l'Institut National Polytechnique de Lorraine, autorise :

Madame BRULLIARD Marie

NANCY BRABOIS
2, AVENUE DE LA
FORET-DE-HAYE
BOITE POSTALE 3
F - 54501
VANDŒUVRE CEDEX

à soutenir devant un jury de l'INSTITUT NATIONAL POLYTECHNIQUE DE LORRAINE, une thèse intitulée :

"Infidélité de transcription et carcinogénèse. Analyse bioinformatique et preuves de concept biolgiques"

en vue de l'obtention du titre de :

DOCTEUR DE L'INSTITUT NATIONAL POLYTECHNIQUE DE LORRAINE

Spécialité : « **Procédés biotechnologiques et alimentaires** »

Fait à Vandoeuvre, le 26 juin 2009
Le Président de l'I.N.P.L.,
F. LAURENT

INSTITUT NATIONAL POLYTECHNIQUE DE LORRAINE - NANCY -

TEL. 33/03.83.59.59.59
FAX. 33/03.83.59.59.55

TITRE DE LA THESE : Infidélité de transcription et carcinogénèse. Analyse bioinformatique et preuves de concept biologiques.

RESUME DE LA THESE

L'un des enjeux de la lutte contre le cancer réside dans la compréhension de l'hétérogénéité de la maladie. Le but de notre travail a été d'explorer l'hétérogénéité des cellules cancéreuses du point de vue de la séquence d'ARN messager. Les ESTs (ou Expressed Sequence Tags) d'origine humaine ont été alignées aux séquences de référence ARNm. Les alignements ont été exploités de manière à mesurer les variations de séquence des ESTs issues de tissus tumoraux ou non tumoraux à chaque position de chaque transcrit. L'analyse statistique mise en place a consisté à identifier les positions pour lesquelles les variations de séquence, *i.e.* substitutions, insertions et délétions, sont différentes entre les ESTs d'origine tumorale et les ESTs d'origine non tumorale. L'étude bioinformatique s'est d'abord concentrée sur 17 transcrits abondamment exprimés avant d'être étendue à l'ensemble du transcriptome. Elle a ensuite été réalisée sur les ESTs murines. Les résultats montrent que l'hétérogénéité des transcrits cancéreux est plus grande que celle des tissus sains. Ainsi, l'infidélité de transcription est augmentée au cours de la carcinogénèse.

Ce résultat bioinformatique a été validé par différentes approches biologiques. Tout d'abord, le clonage puis le séquençage d'un ARN provenant d'une tumeur pulmonaire humaine et présentant une délétion prédite de manière bioinformatique ont été réalisés, et ce, en l'absence de mutation somatique. Ensuite, l'identification par spectrométrie de masse d'un variant protéique issu de la traduction d'un ARN dont le codon stop est substitué en triplet codant a été possible. Enfin, l'intérêt de rechercher dans le sérum de patients cancéreux la présence d'anticorps dirigés contre des protéines issues de la traduction d'ARNm infidèles a été démontré.

Ainsi, l'infidélité de transcription est un phénomène augmenté dans le cancer et responsable d'une partie de l'hétérogénéité des cellules cancéreuses. L'intérêt de cette découverte réside dans les perspectives nouvelles qu'elle offre en termes de compréhension des mécanismes de carcinogénèse et en termes de diagnostic de la maladie.

MOTS CLES : Bioinformatique, expressed sequence tags, cancer.

TITLE : Transcription infidelity and carcinogenesis. Bioinformatical analysis and biological proofs of principle.

ABSTRACT

One of the aim of the fight against cancer is to understand the heterogeneity of cancer cells. The goal of our work has been to explore cancer cell mRNA heterogeneity. ESTs (Expressed Sequence Tags) extracted from normal and cancer tissues have been aligned to mRNA reference sequences. This allowed identification of non-random sequence variations that occurred at statistically significant increased rates in cancer compared to normal libraries. This analysis first focused on 17 abundant transcripts and was next extended to whole human genome, as well as to that of *Mus musculus*. The results show an increase of transcription infidelity events in cancer tissues. Three types of events occur, *i.e.* base substitutions, deletions and insertions.

Bioinformatics results have been validated through different biological methods. First, the cloning and sequencing of mRNA from lung cancer human with a deletion occurring at bioinformatically predicted position in absence of somatic mutation has been achieved. Then, mass spectrometry analysis confirmed the existence of protein variants resulting from translation of mRNA bypassing stop codon. Finally, we showed that transcription infidelity peptides contain specific epitopes of immunoglobulins ; detection of changes in immunoglobulins in patients with cancers opens a novel path toward early stage cancer diagnosis.

This increased transcription infidelity in cancer contributes to the heterogeneity of cancer cells. This finding opens novel perspectives and strategies toward understanding carcinogenesis and diagnostic of the disease.

KEYWORDS : Bioinformatics, expressed sequence tags, cancer.